

RES7001		Research Methodology				L	T	P	J	C
						3	0	0	4	4
Objectives	1. To instruct methods of scientific research. 2. To introduce the concept of design of experiment. 3. To familiarize data collection, analysis, interpretation and report making.									
Expected Outcomes	Selection of an independent research problem and execution, preparation of manuscripts, reports and project proposal for funding agencies									
J-Component	Based on 6 th and 7 th modules									
Modules	Topics					Hours		SLO		
	Common to all Departments									
1	Basic Concept: Importance of research, objectives of research, motivation in research, understanding research and its goal, types of research, research approaches, research methods versus methodology, research and scientific method, research process, criteria of good research.					7		1		
2	Research Ethics: Ethical and moral issues in research, treatment of human subjects and animals in research, copy right laws, authorship issues in publications, intellectual property rights, patent rights, accountability and reproducibility in research.					6		7		
3	Research Design: Need for research design, features of good design, concepts related to research design, basic principle of experimental design and theoretical estimation, design tools.					6		2,3,9		
4	Data Collection: Primary data and secondary data, methods of data collection, important data available for scientists in World Wide Web, reliability of public domain data bases and its implication in research.					6		4		
5	Data Analysis: Data preparation – univariate analysis: frequency tables, bar charts, pie charts, percentages. Bivariate analysis: cross tabulations and chi-square test, hypothesis of association, error analysis.					7		6,10,12		
Department centered										
6	Interpretation and Report writing: Importance of interpretation, techniques of interpretation: precautions in interpretation, significance of report writing, different steps in writing report, layout of the research report, types of reports, oral presentation, writing a good research report.					5		10,13		
7	Publication and Project proposal preparation: Role of scientific journal in research, different ways of citing a research article in manuscripts, Impact factor of journals, importance of citations, <i>h</i> -index, <i>i</i> 10 – index, cite score, plagiarism - software to detect plagiarism, reviewer comments. Funding agencies in India for Science, Engineering and Technology, preparation of a project proposal for funding.					6		8,10		
8	Lecture by experts					2				
	Total Lecture Hours					45				

References:

1. Research Methodology Methods and Techniques, by C R Kothari, (2014) 3rd Edition, New Age International Publishers
2. Research Methodology A step-by-step guide for beginner, by Ranjit Kumar,(2011), SAGE Publication, New Delhi, India
3. Research design and methods a process approach, by Kenneth.S. Borden& Bruce B.Abbott, (2002), Tata Ma-Graw-Hill companies Inc.USA
4. Research Methodology, by Bill Taylor, Gautham Sinha &TaposhGhoshal, (2006) PHI Learning Private Ltd., New Delhi, India.

Approved by Academic Council No.:44

Date:16.03.2017

RES7001 RESEARCH METHODOLOGY

by

Dr.K.KARTHIKEYAN

Department of Maths, SAS, VIT Vellore

MEANING OF RESEARCH

Research is

- A search for knowledge
- A scientific and systematic search for pertinent information on a specific topic
- An art of scientific investigation
- A careful investigation or inquiry especially through search for new facts in any branch of knowledge

- A Systematized effort to gain new knowledge
- A movement from the known to the unknown
- An academic activity and as such the term should be used in a technical sense
- Research comprises defining and redefining problems, formulating hypothesis or suggested solutions; collecting, organising and evaluating data; making deductions and reaching conclusions; and at last carefully testing the conclusions to determine whether they fit the formulating hypothesis.

- An original contribution to the existing stock of knowledge making for its advancement
- The search for knowledge through objective and systematic method of finding solution to a problem is research
- The systematic method consisting of enunciating the problem, formulating a hypothesis, collecting the facts or data, analysing the facts and reaching certain conclusions either in the form of solutions(s) towards the concerned problem or in certain generalisations for some theoretical formulation

OBJECTIVES OF RESEARCH

The purpose of research is to discover answers to questions through the application of scientific procedures

- To gain familiarity with a phenomenon or to achieve new insights into it (studies with this object in view are termed as *exploratory* or *formulative* research studies)
- To portray accurately the characteristics of a particular individual, situation or a group (studies with this object in view are known as *descriptive* research studies)

- To determine the frequency with which something occurs or with which it is associated with something else (studies with this object in view are known as *diagnostic* research studies)
- To test a hypothesis of a causal relationship between variables (such studies are known as *hypothesis-testing* research studies).

MOTIVATIONS IN RESEARCH

The possible motives for doing research may be either one or more of the following:

- Desire to get a research degree along with its consequential benefits;
- Desire to face the challenge in solving the unsolved problems,
- Desire to get intellectual joy of doing some creative work

- Desire to be of service to society
- Desire to get respectability
- Many factors such as directives of government, employment conditions, curiosity about new things, desire to understand causal relationships, social thinking and awakening, and the like may as well motivate (or at times compel) people to perform research operations.

TYPES OF RESEARCH

i. Descriptive vs. Analytical

Descriptive research includes surveys and fact-finding enquiries of different kinds.

The methods of research adopted in conducting descriptive research are survey methods of all kinds, including correlational and comparative methods

Analytical research, the researcher has to use the already available facts or information, and analyse them to make a critical evaluation of the subject.

Descriptive research

To find out the characteristics of a particular entity, process, condition or a grouping

It aims to describe the state of affairs as of present

May involve surveys (also called sample surveys)

May include fact finding enquiries

It is called “Ex post-facto research” in social sciences

In descriptive research, a researcher has no control over the variables that influence the process or phenomena

The outcome is a report as to what happened or what is happening

May include comparative and correlational methods

Descriptive Research Steps

- ↳ Statement of the problem.
 - ↳ Identification of information.
 - ↳ Selection or development of data gathering instruments.
 - ↳ Identification of target population and sample.
 - ↳ Design of information collection procedure.
 - ↳ Collection of information.
 - ↳ Analysis of information.
 - ↳ Generalization and/or predictions.
-

Examples of Descriptive research

What are the brand-preference of customers

Suppose a customer wishes to buy a new smart phone. Which brand are they likely to consider?

It depends on various things which are called variables in research

This may include education level, finance available, purpose for which the phone is bought, ease of usage, features available on a phone etc.

Another example is frequency of online shopping

Some people prefer always to do online shopping as it is convenient, economical, and safe (as items can be returned if customer is not satisfied with it). What are the shopping preferences of college students? This can be a descriptive research

ii). Applied Versus Fundamental

Applied research aims at finding a solution for an immediate problem facing a society or an industrial/business organisation

Fundamental research is mainly concerned with generalisations and with the formulation of a theory

Research aimed at certain conclusions (say, a solution) facing a concrete social or business problem is an example of applied research

Research concerning some natural phenomenon or relating to pure mathematics are examples of fundamental research.

Research studies, concerning human behaviour carried on with a view to make generalisations about human behaviour, are also examples of fundamental research

iii) Quantitative vs. Qualitative

Quantitative research relates to aspects that can be quantified or can be expressed in terms of quantity. It involves the measurement of quantity or amount.

Various available statistical and econometric methods are adopted for analysis in such research which includes correlation, regressions and time series analysis etc.,

Qualitative research is concerned with qualitative phenomena, or more specifically, the aspects related to or involving quality or kind.

Attitude or opinion research i.e., research designed to find out how people feel or what they think about a particular subject or institution is also qualitative research

Qualitative research is particularly significant in the context of behavioural sciences, which aim at discovering the underlying motives of human behaviour.

Qualitative research helps to analyse the various factors that motivate human beings to behave in a certain manner, besides contributing to an understanding of what makes individuals like or dislike a particular thing.

(iv).Conceptual vs. Empirical

Conceptual research is that related to some abstract idea(s) or theory. It is generally used by philosophers and thinkers to develop new concepts or to reinterpret existing ones.

Empirical research relies on experience or observation alone, often without due regard for system and theory. It is data-based research, coming up with conclusions which are capable of being verified by observation or experiment.

We call it as Experimental type of research.

Empirical research is appropriate when proof is sought that certain variables affect other variables in some way. Evidence gathered through experiments or empirical studies is today considered to be the most powerful support possible for a given hypothesis

RESEARCH APPROACHES

Two basic approaches to research

i) *Quantitative approach* ii) *qualitative approach*

Quantitative approach involves the generation of data in quantitative form which can be subjected to rigorous quantitative analysis in a formal and rigid fashion

This approach sub-classified into *inferential, experimental* and *simulation approaches* to research

- Inferential: The purpose of *inferential approach* to research is to form a data base from which to infer characteristics or relationships of population
- *Experimental approach* is characterised by much greater control over the research environment and in this case some variables are manipulated to observe their effect on other variables
- *Simulation approach* involves the construction of an artificial environment within which relevant information and data can be generated. This permits an observation of the dynamic behaviour of a system (or its sub-system) under controlled conditions

QUALITATIVE APPROACH

- *Qualitative approach* to research is concerned with subjective assessment of attitudes, opinions and behaviour
- Research in such a situation is a function of researcher's insights and impressions. Such an approach to research generates results either in non-quantitative form or in the form which are not subjected to rigorous quantitative analysis
- Generally, the techniques of focus group interviews, projective techniques and depth interviews are used

Differences between qualitative and quantitative research

Involves unstructured interviews, observation, and content analysis.

Subjective

Inductive

Little structure

Little manipulation of subjects

Takes a great deal of time to conduct

Little social distance between researcher and subject

Involves experiments, surveys, testing, and structured content analysis, interviews, and observation.

Objective

Deductive

High degree of structure

Some manipulation of subjects

May take little time to conduct

Much social distance between researcher and subject

RESEARCH METHODS VERSUS AND METHODOLOGY

Research Techniques refers to the behaviour and instruments we use in performing research operations such as making observations, recording data, techniques of processing data and like.

Research methods refers to the behaviour and instruments used in selecting and constructing research technique

<i>Type</i>	<i>Methods</i>	<i>Techniques</i>
1. Library Research	(i) Analysis of historical records (ii) Analysis of documents	Recording of notes, Content analysis, Tape and Film listening and analysis. Statistical compilations and manipulations, reference and abstract guides, contents analysis.
2. Field Research	(i) Non-participant direct observation (ii) Participant observation (iii) Mass observation (iv) Mail questionnaire (v) Opinionnaire (vi) Personal interview (vii) Focused interview (viii) Group interview (ix) Telephone survey (x) Case study and life history	Observational behavioural scales, use of score cards, etc. Interactional recording, possible use of tape recorders, photo graphic techniques. Recording mass behaviour, interview using independent observers in public places. Identification of social and economic background of respondents. Use of attitude scales, projective techniques, use of sociometric scales. Interviewer uses a detailed schedule with open and closed questions. Interviewer focuses attention upon a given experience and its effects. Small groups of respondents are interviewed simultaneously. Used as a survey technique for information and for discerning opinion; may also be used as a follow up of questionnaire. Cross sectional collection of data for intensive analysis, longitudinal collection of data of intensive character.
3. Laboratory Research	Small group study of random behaviour, play and role analysis	Use of audio-visual recording devices, use of observers, etc.

RESEARCH METHODS VERSUS AND METHODOLOGY

It seems appropriate at this juncture to explain the difference between research methods and research methodology. *Research methods* may be understood as all those methods/techniques that are used for conduction of research. *Research methods or techniques**, thus, refer to the methods the researchers

*At times, a distinction is also made between research techniques and research methods. *Research techniques* refer to the behaviour and instruments we use in performing research operations such as making observations, recording data, techniques of processing data and the like. *Research methods* refer to the behaviour and instruments used in selecting and constructing research technique. For instance, the difference between methods and techniques of data collection can better be understood from the details given in the following chart—

Type	Methods	Techniques
1. Library Research	(i) Analysis of historical records	Recording of notes, Content analysis, Tape and Film listening and analysis.
	(ii) Analysis of documents	Statistical compilations and manipulations, reference and abstract guides, contents analysis.
2. Field Research	(i) Non-participant direct observation	Observational behavioural scales, use of score cards, etc.
	(ii) Participant observation	Interactional recording, possible use of tape recorders, photo graphic techniques.
	(iii) Mass observation	Recording mass behaviour, interview using independent observers in public places.
	(iv) Mail questionnaire	Identification of social and economic background of respondents.
	(v) Opinionnaire	Use of attitude scales, projective techniques, use of sociometric scales.
	(vi) Personal interview	Interviewer uses a detailed schedule with open and closed questions.
	(vii) Focused interview	Interviewer focuses attention upon a given experience and its effects.
	(viii) Group interview	Small groups of respondents are interviewed simultaneously.
	(ix) Telephone survey	Used as a survey technique for information and for discerning opinion; may also be used as a follow up of questionnaire.
	(x) Case study and life history	Cross sectional collection of data for intensive analysis, longitudinal collection of data of intensive character.
3. Laboratory Research	Small group study of random behaviour, play and role analysis	Use of audio-visual recording devices, use of observers, etc.

From what has been stated above, we can say that methods are more general. It is the methods that generate techniques. However, in practice, the two terms are taken as interchangeable and when we talk of research methods we do, by implication, include research techniques within their compass.

RESEARCH AND SCIENTIFIC METHOD

Research is termed as “An inquiry into the nature of, the reasons for, and the consequences of any particular set of circumstances, whether these circumstances are experimentally controlled or recorded just as they occur.

Research implies the researcher is interested in more than particular results; he/she is interested in the repeatability of the results and in their extension to more complicated and general situations.

The philosophy common to all research methods and techniques, although they may vary considerably from one science to another, is usually called as scientific method

Karl Pearson writes, “The scientific method is one and same in the branches (of science) and that method is the method of all logically trained minds the unity of all sciences consists alone in its methods, not its material; the man who classifies facts of any kind whatever, who sees their mutual relation and describes their sequences, is applying the Scientific Method and is a man of science.”

Scientific methods consist of systematic observation, classification and interpretation of data.

Scientific method is the pursuit of truth as determined by logical considerations.

The ideal of science is to achieve a systematic interrelation of facts.

Scientific method attempts to achieve “this ideal by experimentation, observation, logical arguments from accepted postulates and a combination of these three in varying proportions.”

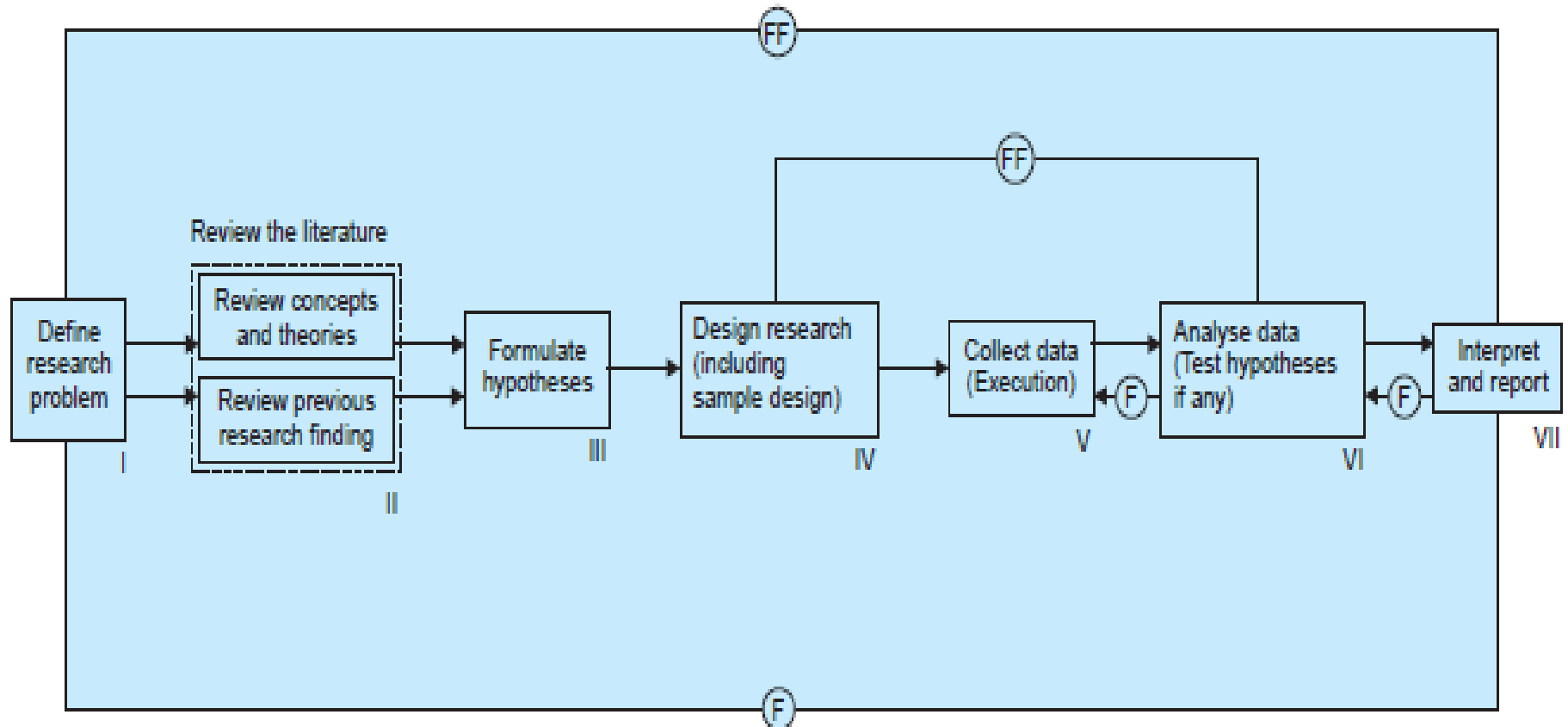
The scientific method is based on certain basic postulates which can be stated as follows

1. It relies on empirical evidence;
2. It utilizes relevant concepts;
3. It is committed to only objective considerations;
4. It presupposes ethical neutrality, i.e., it aims at nothing but making only adequate and correct statements about population objects;
5. It results into probabilistic predictions;
6. Its methodology is made known to all concerned for critical scrutiny are for use in testing the conclusions through replication;
7. It aims at formulating most general axioms or what can be termed as scientific theories

RESEARCH PROCESS

Research process consists of series of actions or steps necessary to effectively carry out research and the desired sequencing of these steps.

RESEARCH PROCESS IN FLOW CHART



Where (F) = feed back (Helps in controlling the sub-system to which it is transmitted)

(FF) = feed forward (Serves the vital function of providing criteria for evaluation)

The following order concerning various steps provides a useful procedural guideline regarding the research process

- (1) Formulating the research problem;
- (2) Extensive literature survey;
- (3) Developing the hypothesis;
- (4) Preparing the research design;
- (5) Determining sample design;
- (6) Collecting the data;
- (7) Execution of the project;
- (8) Analysis of data;
- (9) Hypothesis testing;
- (10) Generalisations and interpretation, and
- 11) Preparation of the report or presentation of the results,

understanding research and its goal

Research depends on the field. A method that works in Chemistry may not be appropriate in Physics or Maths. Research may involve any or all of the following:

Surveys

Interviews

Observation

Experiments

Archival and Historical Data

Qualitative Analysis

Quantitative Analysis

CRITERIA OF GOOD RESEARCH

All types of research work and studies, meet on the common ground of scientific method employed by them

Scientific research satisfy the following criteria:

1. The purpose of the research should be clearly defined and common concepts be used.
2. The research procedure used should be described in sufficient detail to permit another researcher to repeat the research for further advancement, keeping the continuity of what has already been attained.
3. The procedural design of the research should be carefully planned to yield results that are as objective as possible

4. The researcher should report with complete frankness, flaws in procedural design and estimate their effects upon the findings.
5. The analysis of data should be sufficiently adequate to reveal its significance and the methods of analysis used should be appropriate. The validity and reliability of the data should be checked carefully.
6. Conclusions should be confined to those justified by the data of the research and limited to those for which the data provide an adequate basis.
7. Greater confidence in research is warranted if the researcher is experienced, has a good reputation in research and is a person of integrity.

The qualities of a good research

1. *Good research is systematic:* Research is structured with specified steps to be taken in a specified sequence in accordance with the well defined set of rules. Systematic characteristic of the research does not rule out creative thinking but it certainly does reject the use of guessing and intuition in arriving at conclusions.
2. *Good research is logical:* Research is guided by the rules of logical reasoning and the logical process of induction and deduction are of great value in carrying out research. Induction is the process of reasoning from a part to the whole whereas deduction is the process of reasoning from some premise to a conclusion which follows from that very premise.

3. *Good research is empirical:* It implies that research is related basically to one or more aspects of a real situation and deals with concrete data that provides a basis for external validity to research results.
4. *Good research is replicable:* This characteristic allows research results to be verified by replicating the study and thereby building a sound basis for decisions

RES7001 RESEARCH METHODOLOGY

Module-2 – Research Ethics

by

Dr.K.Karthikeyan

Department of Maths,

SAS,VIT Vellore

Ethics

Ethics or moral philosophy is a branch of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior."

Ethics seeks to resolve questions of human **morality** by defining **concepts** such as good and evil, **right and wrong**, virtue and vice, justice and crime

Ethics is about the personal and public judgement as to what is desirable and undesirable, right and wrong and what we 'ought' and 'ought not' to do in areas that are contested.

Morals and Morality

- **Morals** are the actual values (ethical preferences) and norms (rules and principles) that an individual or body accept as guidance for their practices
- A **Morality** is the general name for the package of norms and values that are shared and deployed by a body.

Ethical and moral issues in research

- Research ethics are the set of ethics that govern how scientific and other research is performed at research institutions such as universities, and how it is disseminated(Disperse throughout)
- When most people think of research ethics, they think about issues that arise when research involves human or animal subjects.

Research Ethics

- Research ethics is a world-wide set of principles governing the way any research involving interaction between the researcher and other humans or human tissue or data relating to humans, is designed, managed and conducted.
- In preparing a research project, the dignity, rights, safety and well-being of human participants must at all times be considered, respected and safeguarded

Conflicts of interest

- An institution may include individuals and bodies with diverse values, morals, moralities. There is no consensus amongst ethicists as to the priority of particular theories, rules and principles.
- Indeed within a institution there will be held values, morals, moralities and ethical systems that are incompatible and incommensurable.
- In the absence of a moral and ethical consensus it is a complex task to create an agreed set of mission statements, procedures, policies and training (Gibbins& Reimer 1999, 94-104; 153-160).

Important of Research Ethics

- It is a reflection of respect for those who 'take part' in research
 - It ensures no unreasonable, unsafe or thoughtless demands are made by researchers
 - It ensures sufficient knowledge is shared by all concerned
 - It imposes a common standard in all the above respects
 - It serves as a role-model for others in your area
 - It has become the norm as an expectation for research activity
- a professional requirement for practitioners in some disciplines e.g. psychology
- ... a requirement for access to participants in others

What Projects Need Ethical Approval?

- Human participants
- Use of the 'products' of human participants
- Animal participants
- Work that potentially impacts on human participants

Current ethical principles

- Transparency
- Trust
- Openness
- Honesty
- Respect for other researchers
- **Autonomy** - The freedom to decide what to do. Even when someone has signed a Consent Form, they must be made aware that they are free to withdraw from the study at any time, without giving a reason. They must also be able to request that the data they have given be removed from the study.
- **Indemnity – Security or Protection against loss**

Transparency

- Your results must be transparent to other researchers
- How did you reach the final result?
- What tools and techniques did you use?
- Will the same result be obtained if your research work is replicated elsewhere?
- Do all parties know what, why, how and when?

Trust:

- Can others trust your reported result?
- What is the error margin you have used?

Misconduct in research

- Fabrication or Falsehood
- Fully document results
- Do not falsify data or results
- Ensure repeatability and preserve all data
- Question your own findings
- Do not exaggerate claims or outcomes
- Plagiarism, misquoting, misappropriation
- Attribute honestly the contribution of others
- Colluding in or concealing the misconduct of others

Honesty

- Honesty gives credibility to your work
- People will trust you more if you are honest in your research
- Plagiarism is the offence against honesty
- Claiming someone else's result as your own is dishonesty
- Copying someone else's result in your work is plagiarism
- Even paraphrasing another persons result comes under dishonesty. Both are unethical
- What you have done, how you have done it, what techniques are used, what info you obtained from other sources will give a clear picture to your readers on your capability and work

Research ethics are important for a number of reasons.

- ▶ They promote the aims of research, such as expanding knowledge.
- ▶ They support the values required for collaborative work, such as mutual respect and fairness. This is essential because scientific research depends on collaboration between researchers and groups.
- ▶ They mean that researchers can be held accountable for their actions. Many researchers are supported by public money, and regulations on conflicts of interest, misconduct, and research involving humans or animals are necessary to ensure that money is spent appropriately.
- ▶ They ensure that the public can trust research. For people to support and fund research, they have to be confident in it.
- ▶ They support important social and moral values, such as the principle of doing no harm to others.

Ethical codes cover the following areas

- **Honesty and Integrity**

Report your research honestly, and that this applies to your methods (what you did), your data, your results, and whether you have previously published any of it. You should not make up any data, including extrapolating unreasonably from some of your results, or do anything which could be construed as trying to mislead anyone.

- **Objectivity**

You should aim to avoid bias in any aspect of your research, including design, data analysis, interpretation, and peer review. For example, you should never recommend as a peer reviewer someone you know, or who you have worked with, and you should try to ensure that no groups are inadvertently excluded from your research. This also means that you need to disclose any personal or financial interests that may affect your research.

- **Carefulness**

Take care in carrying out your research to avoid careless mistakes. You should also review your work carefully and critically to ensure that your results are credible. It is also important to keep full records of your research.

- **Openness**

You should always be prepared to share your data and results, along with any new tools that you have developed, when you publish your findings, as this helps to further knowledge and advance science. You should also be open to criticism and new ideas.

- **Respect for Intellectual Property**

You should never plagiarise, or copy, other people's work and try to pass it off as your own. You should always ask for permission before using other people's tools or methods, unpublished data or results.

- **Confidentiality**

You should respect anything that has been provided in confidence. You should also follow guidelines on protection of sensitive information such as patient records.

- **Responsible Publication**

You should publish to advance to state of research and knowledge, and not just to advance your career. This means, in essence, that you should not publish anything that is not new, or that duplicates someone else's work.

- **Legality**

You should always be aware of laws and regulations that govern your work, and be sure that you conform to them.

- **Animal Care**

If you are using animals in your research, you should always be sure that your experiments are both necessary and well-designed. You should also show respect for the animals you are using, and make sure that they are properly cared for.

- **Human Subjects Protection**

If your research involves people, you should make sure that you reduce any possible harm to the minimum, and maximize the benefits both to participants and other people.

Source: Resnick, D. B. (2015) What is Ethics in Research and Why is it Important? List adapted from Shamoo A and Resnik D. 2015. Responsible Conduct of Research, 3rd ed. (New York: Oxford University Press).

ETHICS AND EXPERIMENTS ON ANIMALS

- General thought that it may be necessary to use laboratory animals in some cases in order to create improvements for people, animals or the environment.
- At the same time, the general opinion is that animals have a moral status, and that our treatment of them should be subject to ethical considerations.

Guidelines to handle Animals

- (i) Animals have an intrinsic (belonging naturally) value which must be respected.
- (ii) Animals are sentient creatures with the capacity to feel pain, and the interests of animals must therefore be taken into consideration.
- (iii) Our treatment of animals, including the use of animals in research, is an expression of our attitudes and influences us as moral actors.

The guidelines reflect all these positions, and stipulate principles and Considerations that can be used as tools when balancing between harm and benefit

GUIDELINES

- *Respect for animals' dignity*
- *Responsibility for considering options (Replace)*
- *The principle of proportionality: responsibility for considering and balancing suffering and benefit*
- *Responsibility for considering reducing the number of animals (Reduce)*
- *Responsibility for minimising the risk of suffering and improving animal welfare (Refine)*
- *Responsibility for maintaining biological diversity*

Refine

- Researchers are responsible for assessing the expected effect on laboratory animals.
- Researchers must minimise the risk of suffering and provide good animal welfare. Suffering includes pain, hunger, thirst, malnutrition, abnormal cold or heat, fear, stress, injury, illness and restrictions on the ability to behave normally/naturally.

Replace

- Researchers are responsible for studying whether there are alternatives to experiments on animals. Alternative options must be prioritised if the same knowledge can be acquired without using laboratory animals

Reduce

- Researchers are responsible for considering whether it is possible to reduce the number of animals the experiment plans to use and must only include the number necessary to maintain the scientific quality of the experiments and the relevance of the results

- *Responsibility when intervening in a habitat*
- *Responsibility for openness and sharing of data and material*
- *Requirement of expertise on animals*
- *Requirement of due care*

Copyright

- Copyright refers to laws throughout the modern world that protect the works of various artists (visual artists, writers, musicians) and how those works may be used.
- Copyright is assumed upon creation of the work of art; ie, there is no registration process necessary.

Copyright

- Gives the owner or survivors of estate rights to the work for what is usually 70 years past death of artist.
- Usually treated as a civil matter in courts

Copyright protection

- Automatic upon creation of the work.
- No registration necessary or possible.
- The C copyright symbol is used merely as a reminder to others that this work is original and such a thing as copyright protection exists.

How long does copyright protection last?

- For 70 years beyond the death of the author if the author is known.
- If author is anonymous, for 70 years from first publication of the work.
- Beyond these limitations, the work may be used freely.

Copyright laws

- Transcription is a substantial reproduction of the words spoken, the speaker will own copyright in the words and a separate copyright will apply to the transcription.
- This is of particular relevance to the recording of in-depth interviews and also applies to a recording on tape or video. The person making the recording will own the copyright in the words.
- Copyright can only be transferred in writing and signed by the person making the transfer. This document is called an assignment.
- If researchers wish to publish large extracts from an interview, it is advisable to obtain a transfer of copyright.

Free Use in spite of Copyright

- First, moral rights must still be observed.
- Anyone may make one or a few copies of protected works that have been “released” for personal use. Must be made from a legal copy. Does not apply to software programs or databases. Does apply to music. Generally you have to do the copying yourself
- In the case of a textbook, only limited copying (a chapter, for example) is possible
- There are special provisions in libraries for copying for informational purposes
- Copying / distributing for those with disabilities is permitted.

World Intellectual Property Organization Copyright Treaty (WIPO Treaty)

- created to extend copyright protection for things digital and web-based.
- Includes US and Sweden as well as 184 member states
- Copyright is automatic, and does not require formal registration.
- However, the US still allows statutory damages and attorney's fees for registered works.

What constitutes Copyright Violation?

- 1. Actual Copying (Striking Similarity, and showing of Access and use of that access.)
- 2. Misappropriation
 - A. Not all elements of a work are copyrightable or protectable – for instance, facts, ideas, themes, or content in the public domain.
 - B. Work must be “substantially similar” (not easily defined).
 - C. Formula – Unprotected elements are subtracted, then remaining elements of work are held up to copyrighted work for comparison – are they “substantially similar”?

Authorship Issues in publications

Authorship credit should be based only on:

- (1) Substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data
- (2) Drafting the article or revising it critically for important intellectual content
- (3) Final approval of the version to be published.

Conditions (1), (2), and (3) must all be met.

Acquisition of funding, the collection of data, or general supervision of the research group, by themselves, do not justify authorship

Guidance from the International Committee of Medical Journal Editors (ICMJE)

How to reduce the incidence of authorship problems

People generally lie about authorship in two ways:

- i) putting down names of people who took little or no part in the research (gift authorship, see below)
- ii) Leaving out names of people who did take part (ghost authorship, see below).

Recommended the following three principles to Preventing a problem:

- i) Encourage a culture of ethical authorship
- ii) Start discussing authorship when you plan your research
- iii) Decide authorship before you start each article

Key concepts in authorship

- **Acknowledgements:**

The ICMJE guidelines state: 'All others who contributed to the work who are not authors should be named in the Acknowledgments, and what they did should be described'

- **Appeals:**

You may ask a journal to withdraw your name from a paper if it has been included against your wishes. such cases.

- **Contributorship:**

The ICMJE guidelines recommend that authors should state their contribution to the project: authors should provide a description of what each contributed, and editors should publish that Information

- **Corresponding author:**

The person who receives the reviewers' comments, the proofs, etc. and whose contact details are printed on the article so that readers can request reprints or contact the research group.

- **First and last authors:**

The first named author is held to have made the greatest contribution to the research. Authors have given the last place to a senior team member who contributed expertise and guidance

- **Ghost authors:**

This phrase is used in two ways

It usually refers to professional writers (often paid by commercial sponsors) whose role is not acknowledged.

The term can also be used to describe people who made a significant contribution to a research project but are not listed as authors

- **Gift authors:**

People who are listed as authors but who did not make a significant contribution to the research and therefore do not fulfil the ICMJE criteria. These are often senior officials whose names are added to curry favour.

A colleague whose name is added on the understanding that s/he will do the same for you, regardless of your contribution to his/her research, but simply to swell your publication lists.

- **Number of authors:**

- **Order of authors:**

Source: How to handle authorship disputes: a guide for new researchers

Tim Albert, trainer in medical writing,

Elizabeth Wager, freelance writer and trainer *The COPE Report 2003*

Indication of authorship problems

- Corresponding author seems unable to respond to reviewers' comments
- Changes are made by somebody not on the author list (check Word document properties to see who made the changes but bear in mind there may be an innocent explanation for this, e.g. using a shared computer, or a secretary making changes)
- Document properties show the manuscript was drafted by someone not on the author list or properly acknowledged (but see above)
- Impossibly prolific author e.g. of review articles/opinion pieces (check also for redundant/overlapping publication) (this may be detected by a Medline or Google search using the author's name)

- Several similar review articles/editorials/opinion pieces have been published under different author names (this may be detected by a Medline or Google search using the article title or key words)
- Role missing from list of contributors (e.g. it appears that none of the named authors were responsible for analysing the data or drafting the paper)
- Unfeasibly long or short author list (e.g. a simple case report with a dozen authors or a randomised trial with a single author)
- Industry-funded study with no authors from sponsor company (this may be legitimate, but may also mean deserving authors have been omitted; reviewing the protocol may help determine the role of employees - see Gotzsche et al. and commentary by Wager)

Co-Authorship

Rule of thumb:

- A co-author should have made direct and substantial contributions to the work (not necessarily to the writing).
- Co-authors share responsibility for the scientific integrity of the paper.
- Generally: authors ordered by the amount of their contribution.
- In the Theory community, author list is sometimes alphabetical

Contributions may include:

- Providing key ideas
- Doing the implementation
- Running experiments / collecting data
- Analysing the data
- Writing up the results

Papers typically have 1-4 authors.

Rarely see large author lists as in physics, but is very common in medicine.

Acknowledgments

- People who made contributions that don't merit co-authorship may (sometimes **must**) be acknowledged elsewhere in the paper.
- Not as good as co-authorship, since it doesn't go on a popularity.
- It is good manners, and costs nothing.
- Majority of readers do not read this section. You may put as many names as you wish in the acknowledgement section.

Discuss with Your Advisor

1. What are the authorship conventions in our field?
2. What are the authorship conventions in your lab?
3. Are students prohibited from submitting papers (even if sole-authored) without your approval?
4. Who owns the code/data/manuscript?

See VIT policy on intellectual property.

Intellectual property rights (IPR)

- This is now-a-days very important. There are lot of students who copy videos, music and other resources and post it for free.
- These may be protected by IPR
- These are applicable not only to books and other published work, but also to devices, machines, processes that you make which has a direct application.
- Algorithms are not patentable as such, but if you implement an algorithm in a microchip, device or machine, that product is patentable.

Patent rights

- Patents are granted for applied research. These are usually kept as trade-secrets because others will copy and utilize it for their own benefit.
- Patents can be filed by individuals, institutions, research organisations or companies
- Patent is country-specific.
- If you file a patent in India, other countries could still use your results
- Filing patents in other countries may require huge patent-lawyer costs
- Patents are usually granted after 2 to 4 years

Plagiarism as Defined in the Honor Code

- ▶ “The appropriation of another person’s ideas, processes, results or words without giving appropriate credit.”
 - Presenting the work of another as one’s own.
 - Failing to credit sources used.
 - Attempting to receive credit for work performed by another.
 - Failing to cite ALL resources used, including databases, Internet and other electronic resources.

Why Cite?

- ▶ Citing is part of good research:
 - Connects to previous research.
 - Adds to the collective knowledge.
 - Assists future researchers.
 - Credits others for their ideas.
 - Defines the ideas that are unique.
 - Gives your ideas authority.
- ▶ Citing is about intellectual honesty:
 - Required in academia
 - Essential to retain integrity

Citing for History:

“...the best professional practice for avoiding a charge of plagiarism is always to be **explicit, thorough, and generous in acknowledging one's intellectual debts.**”

--Statement on Standards of Professional Conduct,
American Historical Association

Types of Plagiarism

- ▶ Deliberate Plagiarism: Knowingly using material from a source without proper citation or credit.
- ▶ Partial Plagiarism: Taking pieces of the work (3-5 words) without proper citation.
- ▶ Rewording or Misquoting (Falsification)
- ▶ “Accidental” Plagiarism: Copying instead of paraphrasing; forgetting to place quotation marks correctly; incorrectly citing material.

Avoiding Plagiarism: Strategies

- ▶ Gathering Research Materials
- ▶ Taking Notes
- ▶ Documenting Sources

Gathering Research Materials

- ▶ Expect the research process to take time!
- ▶ Learn the research tools needed for your project.
- ▶ Expect to use library resources---either in print or online. (Research beyond Google!)
- ▶ Allocate time for gathering materials.
- ▶ Leave enough time to carefully read and synthesize your research materials.

Taking Notes

- ▶ Document words that you copy *directly* from a source: do this as you are taking your notes!
- ▶ Jot down the page number and the author/title of the source each time you make a note (even when paraphrasing!)
- ▶ Plagiarism can occur with even *one sentence*.
- ▶ Keep a working bibliography.

Tracking citations and primary sources:

- ▶ Try to trace citations back to the original source; don't depend on a "citation-of-a-citation" in a literature review.
- ▶ If you use someone else's field notes on a primary source, cite the author of the notes, *not* the source.
 - NEVER claim you saw a primary source based on someone else's notes. Bias and translation can misinterpret the source.

Documenting Sources; you must cite...

- ▶ Direct quotes (and be sure to indicate with punctuation/block quotes).
- ▶ Paraphrased passages.
- ▶ Anything that is not *common knowledge*.
 - Common knowledge does NOT include professional terminology; definitions must be cited.
 - Common knowledge is not easily defined; *when in doubt, cite.*
- ▶ Even *unpublished* work must be cited.

Managing Citations

- ▶ Keep a research journal or notebook.
- ▶ Use a bibliographic management tool:
 - EndNote (version X7 for Windows and Mac)
 - Available from Access Services
 - Zotero
 - Mendeley
 - ProCite

Plagiarism

Using someone's work without giving credit or without obtaining permission, where necessary

- Borrowing “just a sentence or two” without attribution is plagiarism.
- Plagiarism is avoided by the citation.
- Copying another person's work
- Downloading articles from internet and using as yours
- Colluding with others for benefit

Citation Etiquette

- Cite other people's work freely and often:
- Avoid antagonizing your reviewers by failing to acknowledge their contributions.
- Demonstrate your mastery of the literature.
- Make new friends. (Scholars love to be cited.)
- Encourage others to cite your work in return.

Citations are good, but stealing citations is not good.

Reviewer Responsibilities

- Do not review manuscripts where you have a personal or professional connection to the author.
- Your friend / relative/ schoolmate.
- Your colleague down the hall.

Know the literature.

- Point out missing citations, especially important contributions.
- Call the editor's attention to any substantial similarity between this manuscript and one already published or currently submitted to another journal.

RES7001 RESEARCH METHODOLOGY

Module-3 – Research Design

Meaning of Research Design

- A research design is the arrangement of conditions for collection and analysis of data that aims to combine relevance to the research purpose with economy in procedure
- The research design is the conceptual structure within which research is conducted; it constitutes the blueprint for the collection, measurement and analysis of data.

The design decisions happen to be in respect of:

- (i) What is the study about?
- (ii) Why is the study being made?
- (iii) Where will the study be carried out?
- (iv) What type of data is required?
- (v) Where can the required data be found?
- (vi) What periods of time will the study include?
- (vii) What will be the sample design?
- (viii) What techniques of data collection will be used?
- (ix) How will the data be analysed?
- (x) In what style will the report be prepared?

Split the overall research design into the following parts:

- (a) *The sampling design* which deals with the method of selecting items to be observed for the given study
- (b) *The observational design* which relates to the conditions under which the observations are to be made;
- (c) *The statistical design* which concerns with the question of how many items are to be observed and how the information and data gathered are to be analysed; and
- (d) *The operational design* which deals with the techniques by which the procedures specified in the sampling, statistical and observational designs can be carried out.

Important features of a research design

- (i) It is a plan that specifies the sources and types of information relevant to the research problem.
- (ii) It is a strategy specifying which approach will be used for gathering and analysing the data.
- (iii) It includes the time and cost budgets since most studies are done under these two constraints.

Research design

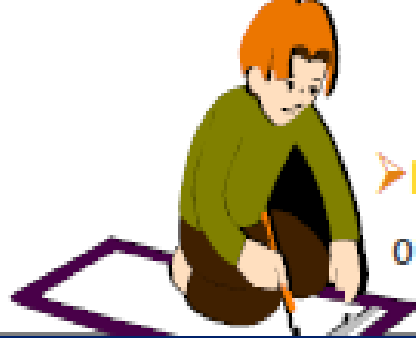
- (a) A clear statement of the research problem
- (b) Procedures and techniques to be used for gathering information
- (c) The population to be studied
- (d) Methods to be used in processing and analysing data.

What is Research Design

It is a **plan** for selecting the sources and types of information used to answer the research question.

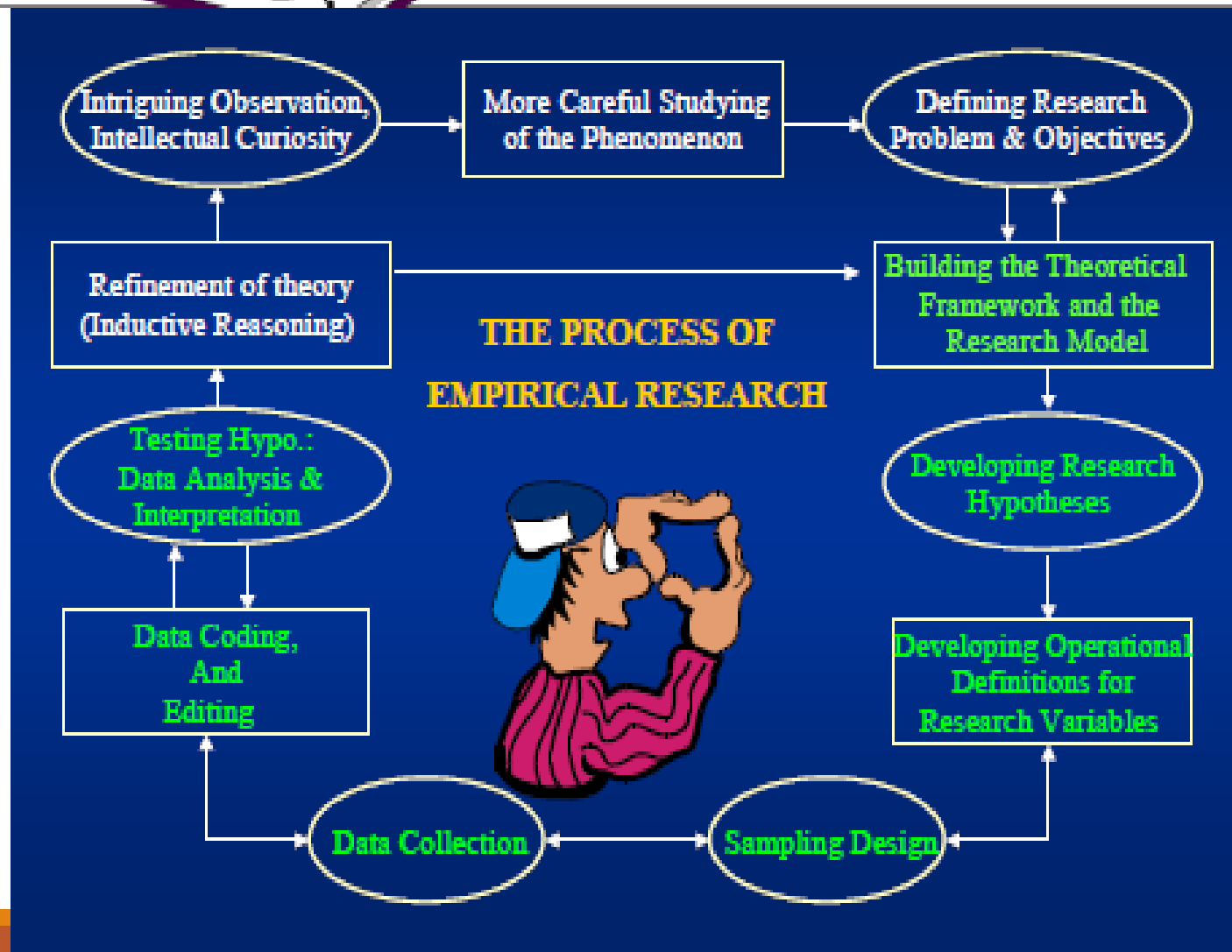
Is a **framework** for specifying the relationships among the study's variables

Is a **blueprint** that outlines each procedure from the hypothesis to the analysis of data.



RESEARCH DESIGN

➤ **RESEARCH DESIGN** refers to the plan, structure, and strategy of research--the blueprint that will guide the research process.



NEED FOR RESEARCH DESIGN

- It facilitates the smooth sailing of the various research operations, thereby making research as efficient as possible yielding maximal information with minimal expenditure of effort, time and money.
- In advance of data collection and analysis for our research project.
- Research design stands for advance planning of the methods to be adopted for collecting the relevant data and the techniques to be used in their analysis, keeping in view the objective of the research and the availability of staff, time and money.

- Research design has great bearing on the reliability of the results arrived at and as such constitutes the firm foundation of the entire edifice of the research work.
- An efficient and appropriate design must be prepared before starting research operations.
- Research design helps the researcher to organize his/her ideas in a form whereby it will be possible for him/her to look for flaws and inadequacies.
- Such a design can even be given to others for their comments and critical evaluation.

FEATURES OF A GOOD DESIGN

- Good design must be flexible, appropriate, efficient, economical and so on.
- Good design minimises bias and maximises the reliability of the data collected and analysed
- Good design gives the smallest experimental error
- Good design yields maximal information and provides an opportunity for considering many different aspects of a problem
- Good design is related to the objective of the research problem and study of the nature of the problem

Factors for research design

- (i) The means of obtaining information
- (ii) The availability and skills of the researcher and his/her staff, if any
- (iii) The objective of the problem to be studied
- (iv) The nature of the problem to be studied
- (v) The availability of time and money for the research

Important concepts relating to research design

i) Dependent and independent variables:

- Variable: A concept which can take on different quantitative values is called a variable.
- Examples: weight, height, income are all examples of variables.
- Qualitative phenomena (or the attributes) are quantified on the basis of the presence or absence of the concerning attribute(s).
- Phenomena which can take on quantitatively different values even in decimal points are called 'continuous variables'.

- If they can only be expressed in integer values, they are 'discrete variables
- Age is an example of continuous variable, but the number of children is an example of non-continuous variable.
- If one variable depends upon or is a consequence of the other variable, it is termed as a dependent variable, and the variable that is antecedent to the dependent variable is termed as an independent variable
- For example, if we say that height depends upon age, then height is a dependent variable and age is an independent variable.
- Height being dependent upon age, height also depends upon the individual's sex, then height is a dependent variable and age and sex are independent variables.

ii) Extraneous variable:

- Independent variables that are not related to the purpose of the study, but may affect the dependent variable are termed as extraneous variables
- For instance the researcher wants to test the hypothesis that there is a relationship between children's gains in social studies achievement and their self-concepts.
- In this case self-concept is an independent variable and social studies achievement is a dependent variable.
- Intelligence may as well affect the social studies achievement, but since it is not related to the purpose of the study undertaken by the researcher, it will be termed as an extraneous variable.

- Whatever effect is noticed on dependent variable as a result of extraneous variable(s) is technically described as an 'experimental error'
- A study must always be so designed that *the effect upon the dependent variable is attributed entirely to the independent variable(s), and not to some extraneous variable or variables.*

iii)Control:

- One important characteristic of a good research design is to minimise the influence or effect of extraneous variable(s).
- The technical term 'control' is used when we design the study minimising the effects of extraneous independent variables.
- In experimental researches, the term 'control' is used to refer to restrain experimental conditions.

iv)Confounded relationship:

- When the dependent variable is not free from the influence of extraneous variable(s), the relationship between the dependent and independent variables is said to be confounded by an extraneous variable(s).

v)Research hypothesis:

- When a prediction or a hypothesised relationship is to be tested by scientific methods, it is termed as research hypothesis.
- The research hypothesis is a predictive statement that relates an independent variable to a dependent variable

vi) Experimental and non-experimental hypothesis-testing research:

- When the purpose of research is to test a research hypothesis, it is termed as hypothesis-testing research. It can be of the experimental design or of the non-experimental design.
- Research in which the independent variable is manipulated is termed 'experimental hypothesis-testing research' and a research in which an independent variable is not manipulated is called 'non-experimental hypothesis-testing research'.
- For instance, suppose a researcher wants to study whether intelligence affects reading ability for a group of students and for this purpose he randomly selects 50 students and tests

their intelligence and reading ability by calculating the coefficient of correlation between the two sets of scores.

- This is an example of non-experimental hypothesis-testing research because herein the independent variable, intelligence, is not manipulated.
- suppose that our researcher randomly selects 50 students from a group of students who are to take a course in statistics and then divides them into two groups by randomly assigning 25 to Group A, the usual studies programme, and 25 to Group B, the special studies programme.
- At the end of the course, he administers a test to each group in order to judge the effectiveness of the training programme on the student's performance-level.

- This is an example of experimental hypothesis-testing research because in this case the independent variable, the type of training programme, is manipulated

vii) Experimental and control groups:

- In an experimental hypothesis-testing research when a group is exposed to usual conditions, it is termed a 'control group', but when the group is exposed to some novel or special condition, it is termed an 'experimental group'.
- In the above illustration, the Group A can be called a control group and the Group B an experimental group.
- If both groups A and B are exposed to special studies programmes, then both groups would be termed 'experimental groups'

viii).Treatments:

- The different conditions under which experimental and control groups are referred as 'Treatments'.
- In the illustration taken above, the two treatments are the usual studies programme and the special studies programme
- To determine through an experiment the comparative impact of three varieties of fertilizers on the yield of wheat, the three varieties of fertilizers treated as three treatments

ix)Experiment:

- The process of examining the truth of a statistical hypothesis, relating to some research problem, is known as an experiment.

For example, we can conduct an experiment to examine the usefulness of a certain newly developed drug.

Two types of experiments : i)absolute experiment ii)comparative experiment.

To determine the impact of a fertilizer on the yield of a crop, it is a case of absolute experiment

To determine the impact of one fertilizer as compared to the impact of some other fertilizer, termed as a comparative experiment

x)Experimental unit(s)

The pre-determined plots or the blocks, where different treatments are used, are known as experimental units.

DIFFERENT RESEARCH DESIGNS

Different research designs are

- Research Design in case of Exploratory research studies
- Research Design in case of Descriptive and diagnostic research studies
- Research Design in case of Hypothesis-testing research studies.

Research design in case of Exploratory research studies

- Exploratory research studies are termed as formulative research studies.
- The main purpose of this studies is that of formulating a problem for more precise investigation or of developing the working hypotheses from an operational point of view.
- The major emphasis in this studies is on the discovery of ideas and insights.

The following three methods in the context of research design for such studies are

(i) The survey of concerning literature;

(ii) The experience survey

(iii) The analysis of 'insight-stimulating' examples.

i)The survey of concerning literature

- The survey of concerning literature happens to be fruitful method of formulating precisely the research problem or developing hypothesis.
- Hypotheses stated by earlier workers may be reviewed and their usefulness be evaluated as a basis for further research.
- It may be considered whether the already stated hypotheses suggest new hypothesis. The researcher should review and build upon the work already done by others.

ii) The experience survey

- An experience survey may enable the researcher to define the problem more concisely and help in the formulation of the research hypothesis.
- This survey may as well provide information about the practical possibilities for doing different types of research.

- iii) *Analysis of ‘insight-stimulating’ examples* is a fruitful method for suggesting hypotheses for research.
- This method consists of the intensive study of selected instances of the phenomenon
- For this purpose the existing records, if any, may be examined, the unstructured interviewing may take place, or some other approach may be adopted.
- Attitude of the investigator, the intensity of the study and the ability of the researcher to draw together diverse information into a unified interpretation are the main features

Research design in case of descriptive and diagnostic research studies

- Descriptive research studies are those studies which are concerned with describing the characteristics of a particular individual, or of a group
- Diagnostic research studies determine frequency with which something occurs or its association with something else.
- The research design must make enough provision for protection against bias and must maximise reliability, with due concern for the economical completion of the research study.

The design must be rigid and not flexible and focus on the following:

- (a) Formulating the objective of the study (what the study is about and why is it being made?)
- (b) Designing the methods of data collection (what techniques of gathering data will be adopted?)
- (c) Selecting the sample (how much material will be needed?)
- (d) Collecting the data (where can the required data be found and with what time period should the data be related?)
- (e) Processing and analysing the data.
- (f) Reporting the findings.

Difference between research designs in respect of the above two types of research studies

<i>Research Design</i>	<i>Type of study</i>	
	<i>Exploratory of Formulative</i>	<i>Descriptive/Diagnostic</i>
Overall design	Flexible design (design must provide opportunity for considering different aspects of the problem)	Rigid design (design must make enough provision for protection against bias and must maximise reliability)
(i) Sampling design	Non-probability sampling design (purposive or judgement sampling)	Probability sampling design (random sampling)
(ii) Statistical design	No pre-planned design for analysis	Pre-planned design for analysis
(iii) Observational design	Unstructured instruments for collection of data	Structured or well thought out instruments for collection of data
(iv) Operational design	No fixed decisions about the operational procedures	Advanced decisions about operational procedures.

Research design in case of hypothesis-testing research studies

- Hypothesis-testing research studies (generally known as experimental studies) are those where the researcher tests the hypotheses of causal relationships between variables.
- This studies require procedures that will not only reduce bias and increase reliability, but will permit drawing inferences about causality. Usually experiments meet this requirement.
- When we talk of research design in such studies, we often mean the design of experiments.

BASIC PRINCIPLES OF EXPERIMENTAL DESIGNS

Professor Fisher has enumerated three principles of experimental designs:

i) Principle of Replication

- The experiment should be repeated more than once.
- Each treatment is applied in many experimental units instead of one.
- Due to this the statistical accuracy of the experiments is increased.

- To examine the effect of two varieties of rice we divide the field in to two parts and grow one variety in one part and the other variety in the other part and then compare the yield of the two parts and draw conclusion on that basis.
- To apply the principle of replication to this experiment, then we first divide the field into several parts, grow one variety in half of these parts and the other variety in the remaining parts.
- Collect the data of yield of the two varieties and draw conclusion by comparing the same which is more reliable in comparison to the conclusion we draw without applying the principle of replication.

ii) Principle of Randomization

- *Principle of Randomization* provides protection, when we conduct an experiment, against the effect of extraneous factors by randomization.
- For instance, if we grow one variety of rice, say, in the first half of the parts of a field and the other variety is grown in the other half, then it is just possible that the soil fertility may be different in the first half in comparison to the other half.
- Results would not be realistic due to this.

- For this, we assign the variety of rice to be grown in different parts of the field on the basis of some random sampling technique we may apply randomization principle and protect ourselves against the effects of the extraneous factors (soil fertility differences in the given case).
- Through the application of the principle of randomization, we can have a better estimate of the experimental error

Principle of Local Control

- Divide the field into several homogeneous parts, known as blocks, and then each such block is divided into parts equal to the number of treatments and the treatments are randomly assigned to these parts of a block
- Blocks are the levels at which hold an extraneous factor fixed, so that we can measure its contribution to the total variability of the data by means of a two-way analysis of variance.
- Through the principle of local control we can eliminate the variability due to extraneous factor(s) from the experimental error.

Experimental Design

Experimental design classified in to two categories

- i) Informal experimental design
- ii) Formal experimental design

- i) Informal experimental designs are those designs that normally use a less sophisticated form of analysis based on differences in magnitudes
- ii) Formal experimental designs offer relatively more control and use precise statistical procedures for analysis

(a) Informal experimental designs:

(i) Before-and-after without control design.

(ii) After-only with control design.

(iii) Before-and-after with control design.

(b) Formal experimental designs:

(i) Completely randomized design (C.R. Design).

(ii) Randomized block design (R.B. Design).

(iii) Latin square design (L.S. Design).

(iv) Factorial designs.

1. Before-and-after without control design:

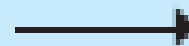
- In such a design a single test group or area is selected and the dependent variable is measured before the introduction of the treatment.
- The treatment is then introduced and the dependent variable is measured again after the treatment has been introduced.
- The effect of the treatment would be equal to the level of the phenomenon after the treatment minus the level of the phenomenon before the treatment.

Test area:

Level of phenomenon
before treatment (X)

Treatment
introduced

Level of phenomenon
after treatment (Y)



$$\text{Treatment Effect} = (Y) - (X)$$

The main difficulty of such a design is that with the passage of time considerable extraneous variations may be there in its treatment effect.

2. After-only with control design:

- In this design two groups or areas (test area and control area) are selected and the treatment is introduced into the test area only.
- The dependent variable is then measured in both the areas at the same time. Treatment impact is assessed by subtracting the value of the dependent variable in the control area from its value in the test area.

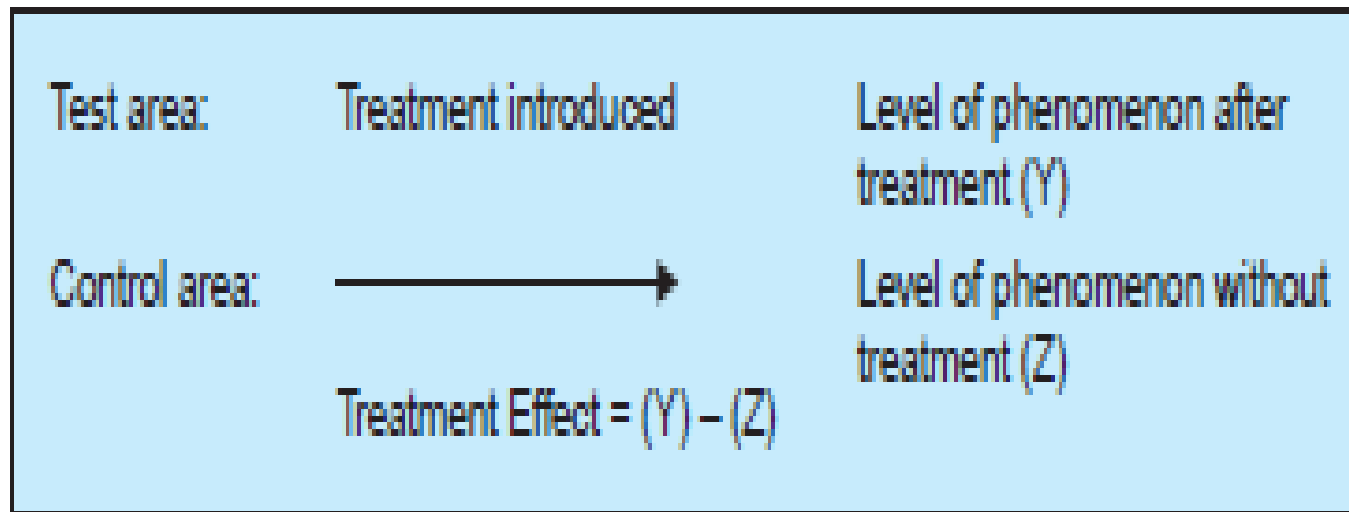


Figure 1.1

The basic assumption in such a design is that the two areas are identical with respect to their behaviour towards the phenomenon considered.

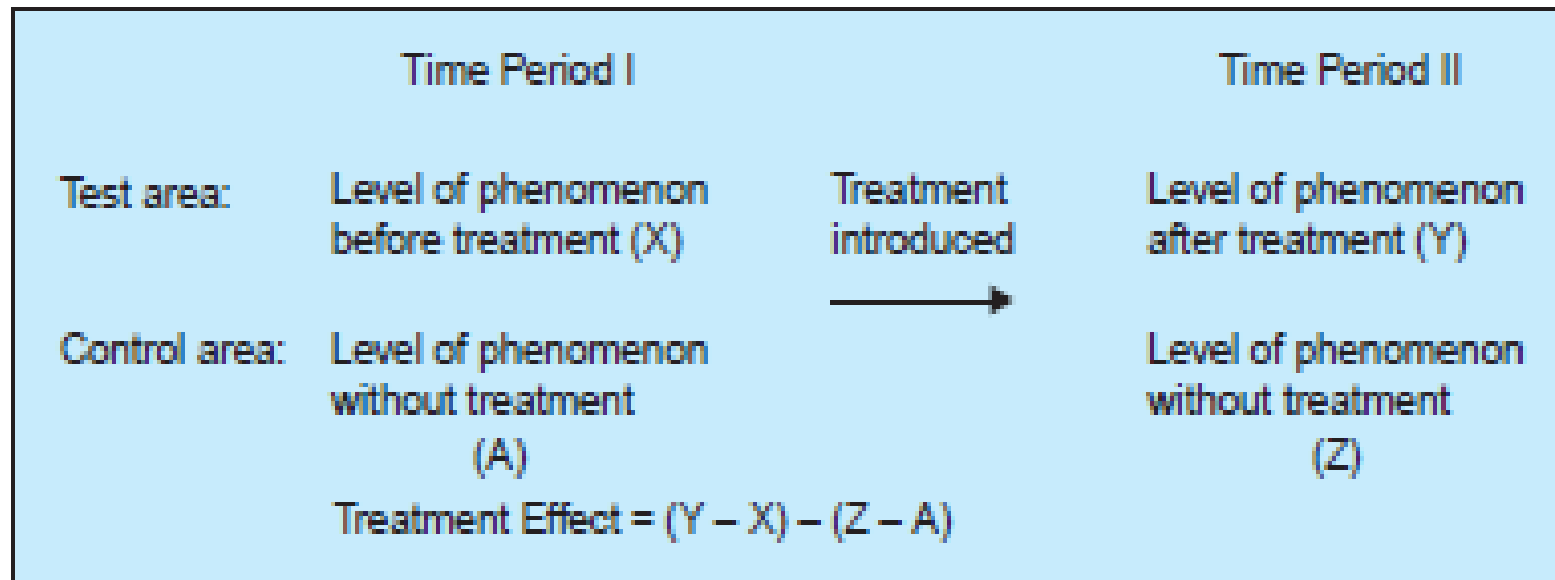
If this assumption is not true, there is the possibility of extraneous variation entering into the treatment effect.

However, data can be collected in such a design without the introduction of problems with the passage of time.

In this respect the design is superior to before-and-after without control design.

3. Before-and-after with control design:

- Two areas are selected and the dependent variable is measured in both the areas for an identical time-period before the treatment.
- The treatment is introduced into the test area only, and the dependent variable is measured in both for an identical time-period after the introduction of the treatment.
- The treatment effect is determined by subtracting the change in the dependent variable in the control area from the change in the dependent variable in test area.



This design is superior to the above two designs since it avoids extraneous variation resulting both from the passage of time and from non-comparability of the test and control areas.

Completely randomized design (C.R. design):

CRD Involves two principles

- i) The principle of replication
- ii) The principle of randomization of experimental designs.

The essential characteristic of the design is that subjects are randomly assigned to experimental treatments

- For example, if we have 10 subjects and if we wish to test 5 under treatment A and 5 under treatment B, the randomization process gives every possible group of 5 subjects selected from a set of 10 an equal opportunity of being assigned to treatment A and treatment B.
- One-way analysis of variance is used to analyse such a design.
- Even unequal replications can also work in this design. Such a design is used when experimental areas happen to be homogeneous.

Two-group simple randomized design:

In a two-group simple randomized design,

First define population then from the population a sample is selected randomly.

Further, requirement of this design is that items, after being selected randomly from the population, be randomly assigned to the experimental and control groups (Such random assignment of items to two groups is technically described as principle of randomization).

This design yields two groups as representatives of the population.

Diagram form for this design

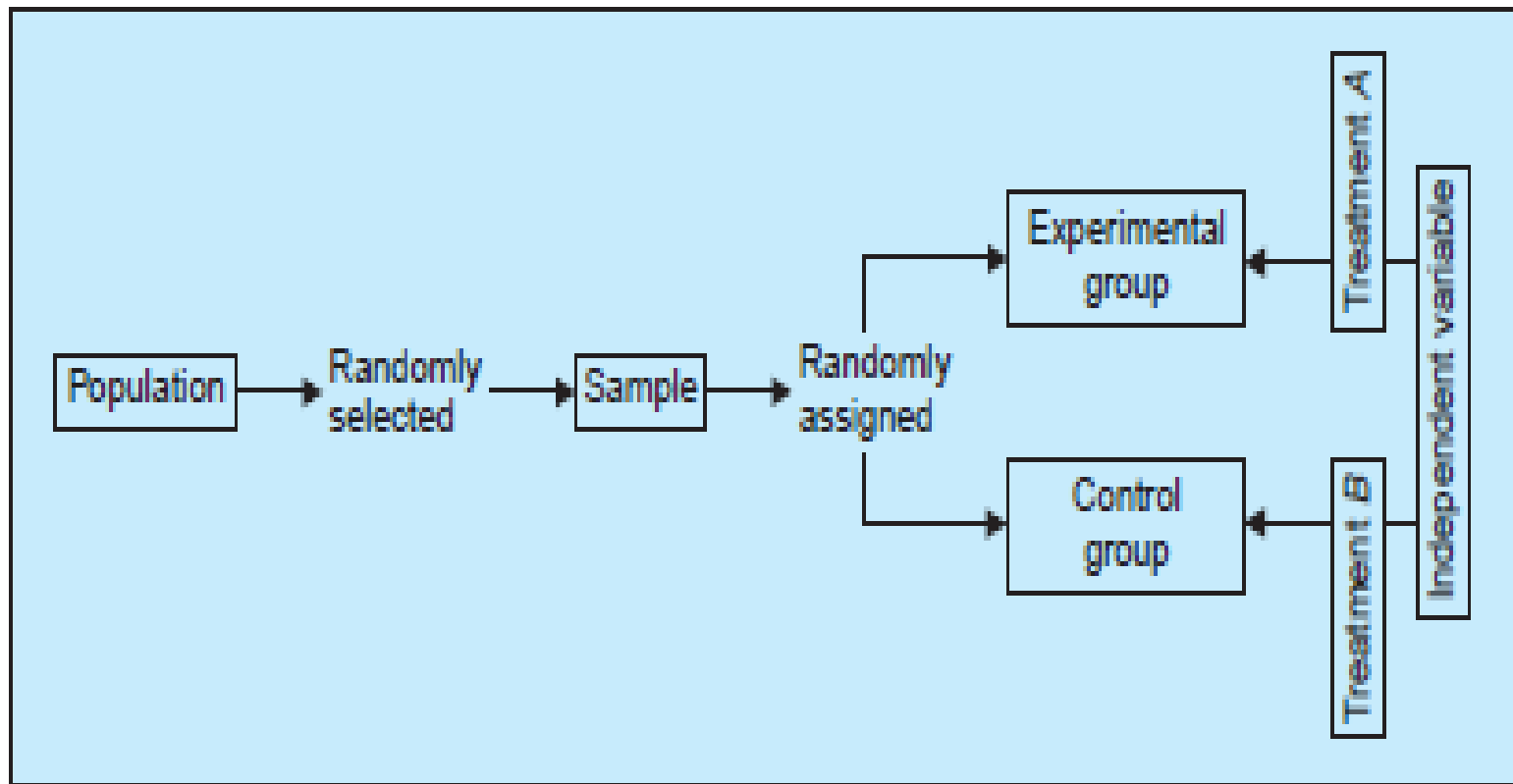


Illustration:

Suppose the researcher wants to compare two groups of students who have been randomly selected and randomly assigned.

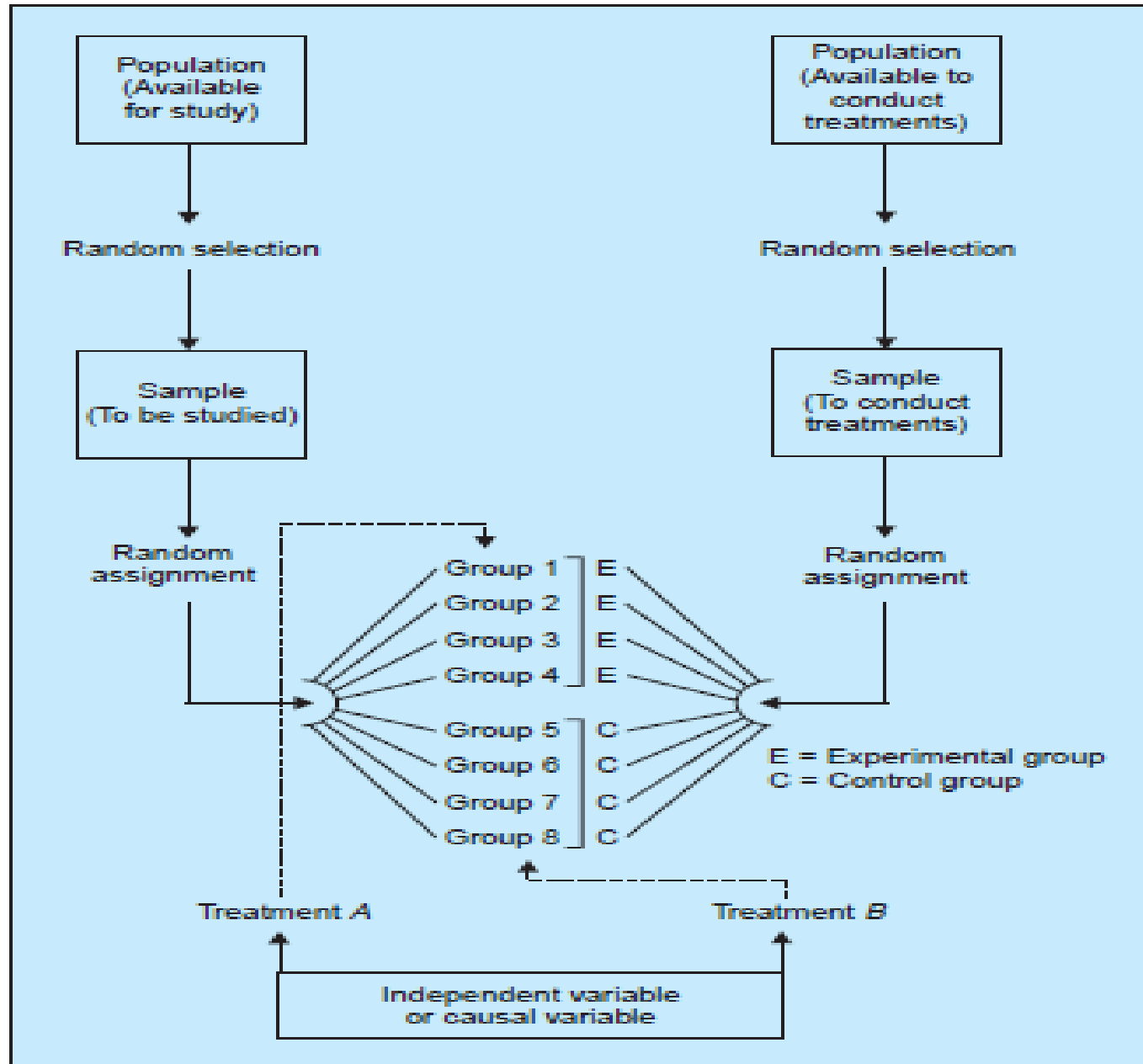
Two different treatments: The usual training and the specialized training are being given to the two groups. The researcher hypothesises greater gains for the group receiving specialised training.

To determine this, researcher tests each group before and after the training, and then compares the amount of gain for the two groups to accept or reject his hypothesis.

This is an illustration of the two-groups randomized design, wherein individual differences among students are being randomized.

But this does not control the differential effects of the extraneous independent variables (in this case, the individual differences among those conducting the training programme).

Random replication design:



It is clear from the diagram that there are two populations in the replication design. The sample is taken randomly from the population available for study and is randomly assigned say, four experimental and four control groups.

Similarly, sample is taken randomly from the population available to conduct experiments (because of the eight groups eight such individuals be selected) and the eight individuals so selected should be randomly assigned to the eight groups.

Variables relating to both population characteristics are assumed to be randomly distributed among the two groups.

Thus, this random replication design is an extension of the two-group simple randomized design.

COMPLETELY RANDOMIZED DESIGN

- Let us suppose that we compare h treatments (h different manures) and there are n plots available for the experiment.
- Let i th treatment be replicated n_i times so that $n_1 + n_2 + n_3 + \dots + n_h = n$
- The plots to which the different treatments are to be given are found by the following randomization principle
- The plots are numbered from 1 to n serially n identical cards are taken, numbered from 1 to n and shuffled thoroughly

- The numbers on the first n_1 cards drawn randomly give the numbers of the plots to which the first treatment is given
- The numbers on the next n_2 cards drawn randomly give the numbers of the plots to which the second treatment is given and so on
- The design is called a completely randomized design which is used when the plots are homogeneous or the pattern of heterogeneity of the plots is unknown.

Randomised Block Design(R.B.D)

- Let us consider an agricultural experiment using which we wish to test the effect of 'k' fertilizing treatments on the yield of a crop.
- We assume that we know some information about the soil fertility of the plots
- Then we divide the plots in to h blocks according to the soil fertility, each block containing k plots
- Thus the plots in each block will be of homogeneous fertility as far as possible

- Within each block, the k treatments are given to h blocks in a perfectly random manner, such that each treatment occurs only once in any block.
- The same k treatments are repeated from block to block. This design is called Randomised Block Design

Latin Square Design(L.S.D)

- Consider an agricultural experiment in which n^2 plots are taken and arranged in the form of an $n \times n$ square such that the plots in each row will be homogeneous as far as possible with respect to one factor of classification, say, soil fertility.
- The plots in each column will be homogeneous as far as possible with respect to another factor of classification say, seed quality
- n treatments are given to these plots such that each treatment occurs only once in each row and only once in each column. This design is called L.S.D.

FACTORIAL DESIGN

The yield of a chemical process may be affected by several factors such as the levels of pressure, temperature, rate of agitation, and proportion of reactants etc.,

The factorial experiments are useful in experimental situations which require the examination of the effects of varying two or more factors.

In a complete exploration of such a situation, it is not sufficient to vary one factor at a time, but that all combinations of the different factor levels must be examine in order to elucidate the effect of each factor.

Let us consider two fertilizers, say, Potash(K) and Nitrogen(N)

let us suppose that there are p different varieties of Potash and q different varieties of Nitrogen. p and q are called as the levels of the factors potash and nitrogen respectively

To find the effectiveness of various treatments, ie different levels of Potash or Nitrogen we might conduct two simple experiments one for Potash and the other for Nitrogen.

A series of experiments in which only one factor is varied at a time would be both lengthy and costly.

These experiments do not give us any information regarding the dependence or independence of one factor on the other. (Do not tell us anything about the intersection effect (NK)).

To investigate the variations in several factors simultaneously by conducting the above experiment as a $p \times q$ factorial experiment where p and q are the levels of various factors

If the levels of various factors are equal then s^n factorial experiment means an experiments with n factors, each at s levels where n is any positive integer greater than or equal to 2. Ex: 2^3 factorial exp. means an experiment with 3 factors at 2 levels each

Factorial designs

- Factorial designs are used in experiments where the effects of varying more than one factor are to be determined.
- They are important in several economic and social phenomena where usually a large number of factors affect a particular problem.
- Factorial designs can be of two types:
 - i) Simple factorial designs
 - ii) Complex factorial designs.

Simple factorial designs (or) Two-factor-factorial design

In simple factorial designs, we consider the effects of varying two factors on the dependent variable

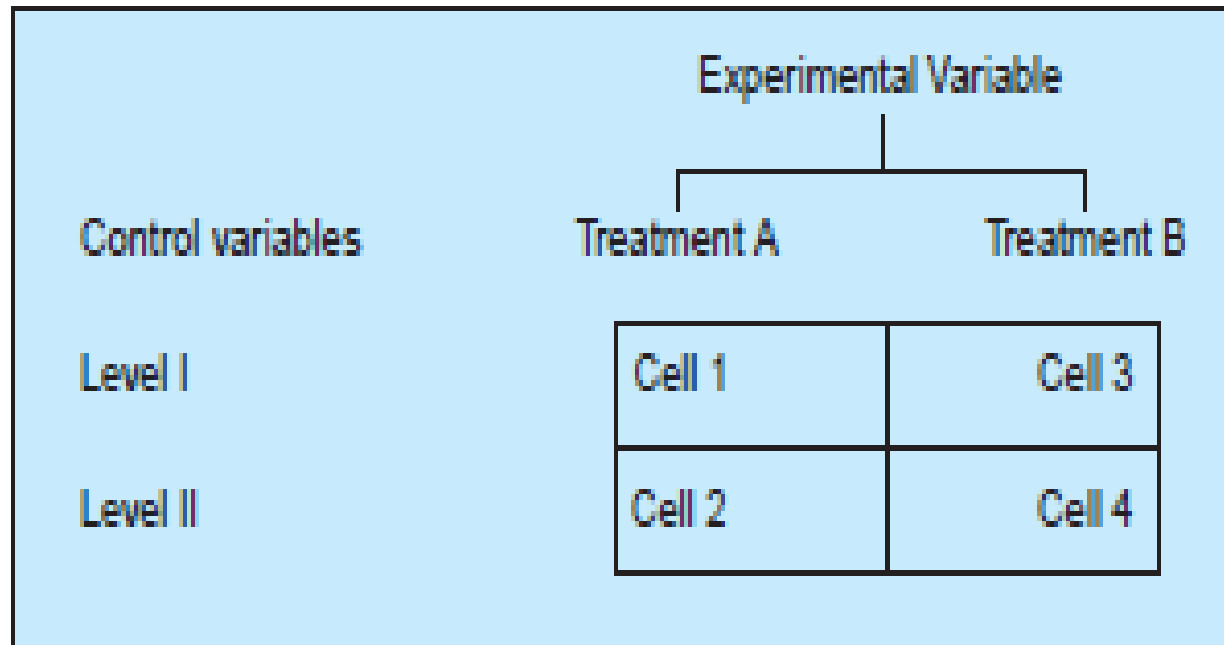
complex factorial design: (or) Multifactor-factorial design

When an experiment is done with more than two factors, we call as complex factorial designs

Simple factorial design be a 2×2

Graphical representation of 2×2 simple factorial design

2×2 SIMPLE FACTORIAL DESIGN



Dr. K. Karthikeyan, SAS

- In this design the extraneous variable to be controlled by homogeneity is called the control variable and the independent variable, which is manipulated, is called the experimental variable.
- There are two treatments of the experimental variable and two levels of the control variable.
- There are four cells into which the sample is divided. Each of the four combinations would provide one treatment or experimental condition.
- The means for different cells be obtained along with the means for different rows and columns.

- Means of different cells represent the mean scores for the dependent variable.
- The column means in the design are the main effect for treatments without taking into account any differential effect due to the level of the control variable.
- The row means in the design are the main effects for levels without regard to treatment.
- By this design we can study the main effects of treatments as well as the main effects of levels.
- An additional merit of this design is that one can examine the interaction between treatments and levels, through which one may say whether the treatment and levels are independent of each other or they are not so.

Interaction effect between treatments and levels and data in case of two (2×2) simple factorial studies

STUDY I DATA

		Training		Row Mean
		Treatment A	Treatment B	
Control (Intelligence)	Level I (Low)	15.5	23.3	19.4
	Level II (High)	35.8	30.2	33.0
	Column mean	25.6	26.7	

STUDY II DATA

		Training		Row Mean
		Treatment A	Treatment B	
Control (Intelligence)	Level I (Low)	10.4	20.6	15.5
	Level II (High)	30.6	40.4	35.5
	Column mean	20.5	30.5	

Fig. 3.9

All the above figures (the study I data and the study II data) represent the respective means. Graphically, these can be represented as shown in Fig. 3.10.

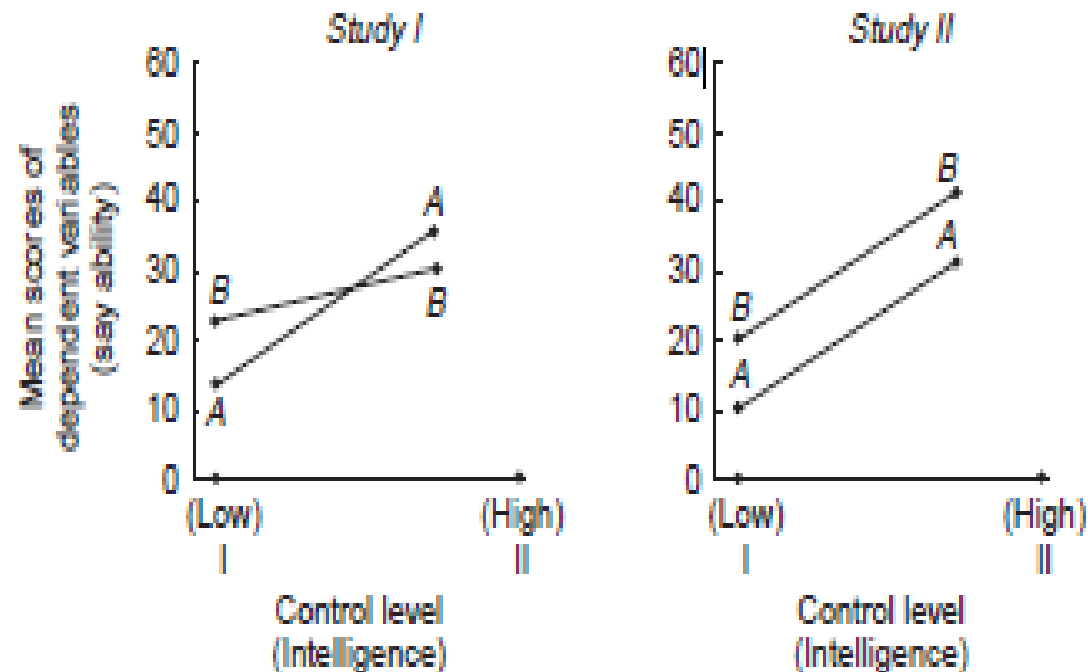


Fig. 3.10

- The graph relating to Study I indicates that there is an interaction between the treatment and the level which, in other words, means that the treatment and the level are not independent of each other.
- The graph relating to Study II shows that there is no interaction effect which means that treatment and level in this study are relatively independent of each other.
- The 2×2 design need not be restricted in the manner as explained above i.e., having one experimental variable and one control variable, but it may also be of the type having two experimental variables or two control variables.

Complex factorial designs

A design which considers three or more independent variables simultaneously is called a complex factorial design.

In case of three factors with one experimental variable having two treatments and two control variables, each one of which having two levels, the design will be termed $2 \times 2 \times 2$ complex factorial design which will contain a total of eight cells

In Fig. 3.14 a pictorial presentation is given of the design shown below.

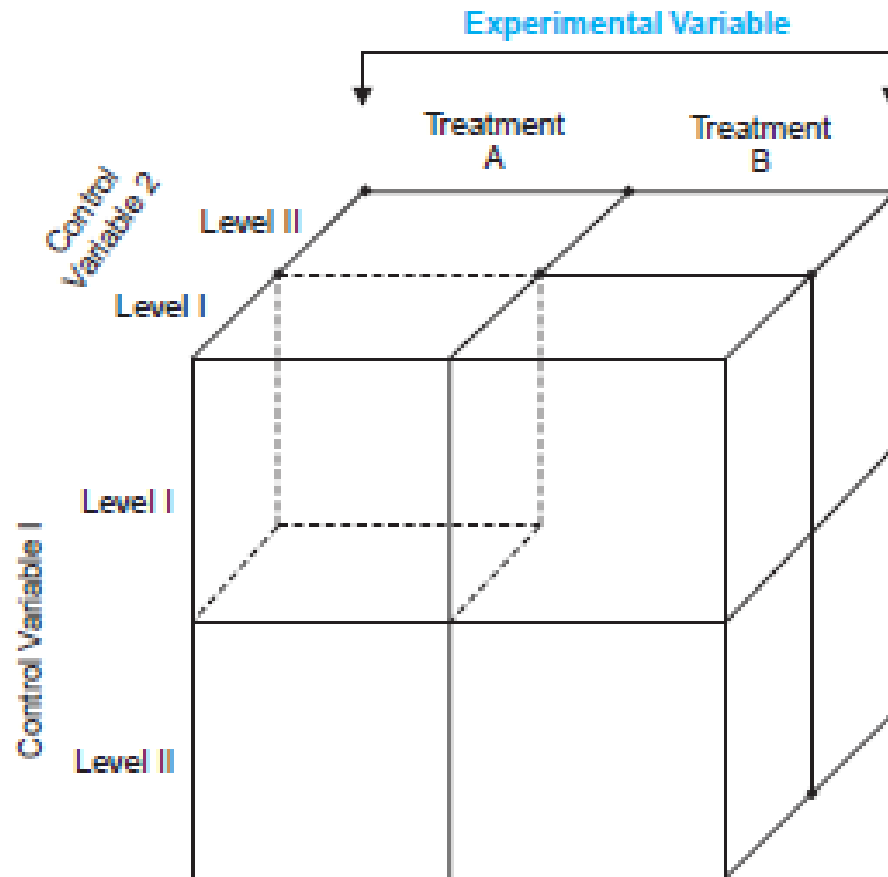


Fig. 3.14

The dotted line cell in the diagram corresponds to Cell 1 of the above stated $2 \times 2 \times 2$ design and is for Treatment A, level I of the control variable 1, and level I of the control variable 2.

From this design it is possible to determine the main effects for three variables i.e., one experimental and two control variables.

The researcher can also determine the interactions between each possible pair of variables (such interactions are called 'First Order interactions') and interaction between variable taken in triplets (such interactions are called Second Order interactions).

In case of a $2 \times 2 \times 2$ design, the further given first order interactions are possible:

Experimental variable with control variable 1

(EV \times CV 1);

Experimental variable with control variable 2

(EV \times CV 2);

Control variable 1 with control variable 2

(CV1 \times CV2);

There will be one second order interaction as well in the given design (it is between all the three variables (EV \times CV1 \times CV2)).

To determine the main effects for the experimental variable, the researcher must necessarily compare the combined mean of data in cells 1, 2, 3 and 4 for Treatment A with the combined mean of data in cells 5, 6, 7 and 8 for Treatment B.

In this way the main effect for experimental variable, independent of control variable 1 and variable 2, is obtained.

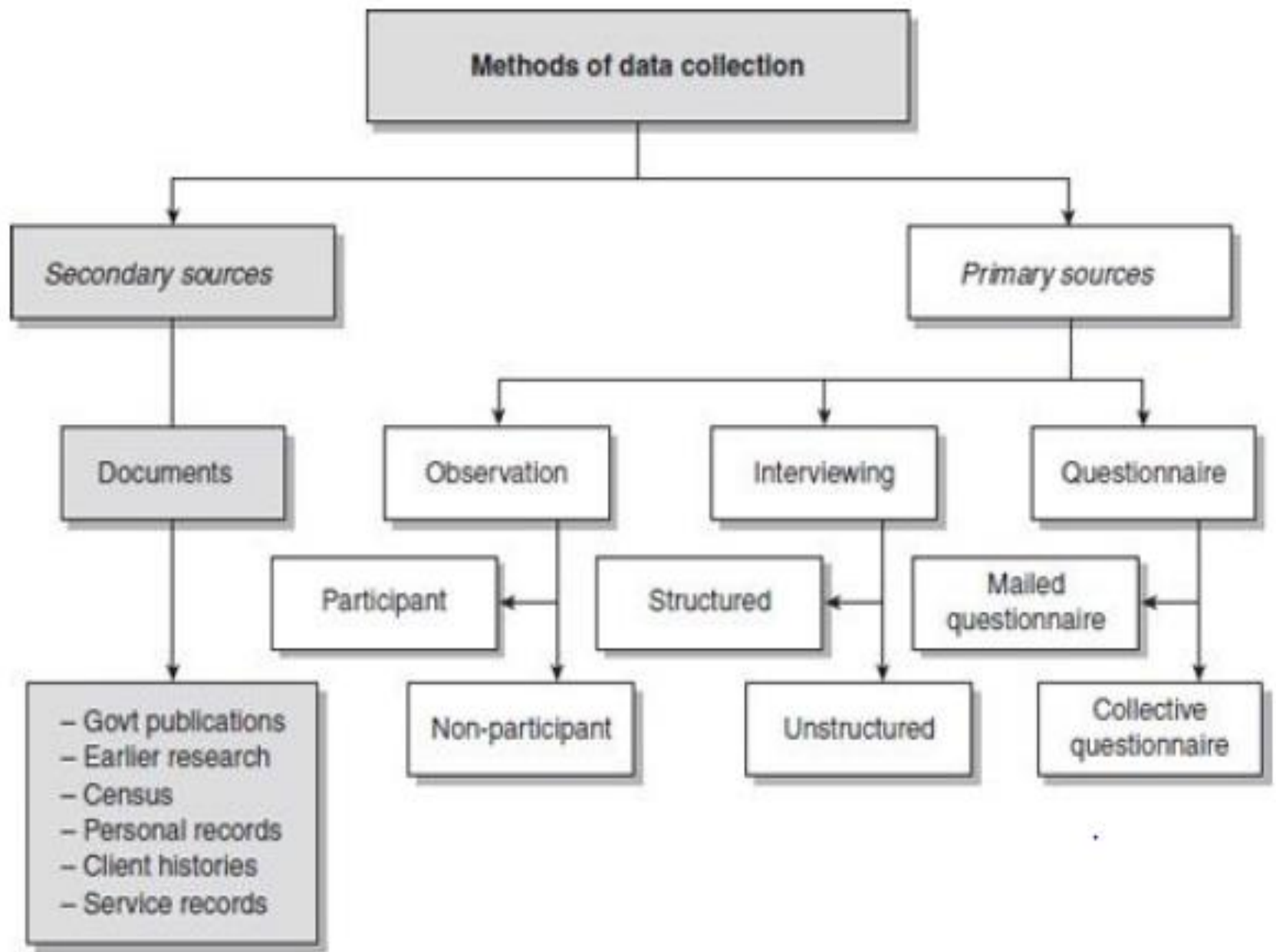
The main effect for control variable 1, independent of experimental variable and control variable 2, is obtained

We compare the combined mean of data in cells 1, 3, 5 and 7 with the combined mean of data in cells 2, 4, 6 and 8 of our $2 \times 2 \times 2$ factorial design and determine the main effect for the control variable 2 independent of experimental variable and control variable 1, if the combined mean of data in cells 1, 2, 5 and 6 are compared with the combined mean of data in cells 3, 4, 7 and 8.

DATA COLLECTION

Module -4

- Data collection begins after research problem and research design are identified
- Two types of data: i) Primary data
ii) Secondary data
- The *primary data* are those which are collected afresh and for the first time, and thus happen to be original in character.
- The *secondary data*, are those which have already been collected by someone else and which have already been passed through the statistical process.



Methods of data collection

- Examples of primary sources include finding out first-hand the attitudes of a community towards health services, ascertaining the health needs of a community, evaluating a social programme, determining the job satisfaction of the employees of an organization, and ascertaining the quality of service provided by a worker are examples of information collected from primary sources.
- Examples of secondary data includes the use of census data to obtain information on the age–sex structure of a population, the use of hospital records to find out the morbidity and mortality patterns of a community, the use of an organization's records to ascertain its activities, and the collection of data from sources such as articles, journals, magazines, books and periodicals to obtain historical and other types of information.
- Primary sources provide first-hand information and secondary sources provide second-hand data.

The choice of a method to collect primary data depends upon the purpose of the study, the resources available and the skills of the researcher.

In selecting a method of data collection, the socio-economic demographic characteristics of the study population play an important role:

Should know as much about characteristics such as educational level, age structure, socio-economic status and ethnic background, the study population's interest in, and attitude towards, participation in the study.

Some populations, for a number of reasons, may not feel either at ease with a particular method of data collection (such as being interviewed) or comfortable with expressing opinions in a questionnaire.

Several methods of collecting primary data, in surveys and descriptive researches are:

- (i) Observation method
 - (ii) Interview method,
 - (iii) Through questionnaires,
 - (iv) Through schedules,
 - (v) Other methods
- (a) Warranty cards; (b) Distributor audits; (c) Pantry audits; (d) Consumer panels; (e) Using mechanical devices; (f) Through projective techniques; (g) Depth interviews, (h) Content analysis.

Observation Method

- Observation method used in specially in studies relating to behavioural sciences.
- The information is sought by way of investigator's own direct observation without asking from the respondent.
- For instance, in a study relating to consumer behaviour, the investigator instead of asking the brand of wrist watch used by the respondent, may himself look at the watch.

Advantages of Observation Method

- i) The subjective bias is eliminated, if observation is done accurately.
- ii) The information obtained under this method relates to what is currently happening; it is not complicated by either the past behaviour or future intentions or attitudes.
- iii) This method is independent of respondents willingness to respond and as such is relatively less demanding of active cooperation on the part of respondents as happens to be the case in the interview or the questionnaire method.

- This method is particularly suitable in studies which deal with subjects (i.e., respondents) who are not capable of giving verbal reports of their feelings for one reason or the other
- Limitations of observation method
 - i) It is an expensive method.
 - ii) The information provided by this method is very limited.
 - iii) Sometimes unforeseen factors may interfere with the observational task. At times, the fact that some people are rarely accessible to direct observation creates obstacle for this method to collect data effectively.

- Two types of observation:
 1. Participant observation;
 2. Non-participant observation

Participant observation is when you, as a researcher, participate in the activities of the group being observed in the same manner as its members, with or without their knowing that they are being observed.

For example, you might want to examine the reactions of the general population towards people in wheel chairs. You can study their reactions by sitting in a wheelchair yourself.

Merits of the Participant type of observation

- (i) The researcher is enabled to record the natural behaviour of the group.
- (ii) The researcher can even gather information which could not easily be obtained if he observes in a disinterested fashion.
- (iii) The researcher can even verify the truth of statements made by informants in the context of a questionnaire or a schedule.

Non-Participant observation

when you, as a researcher, do not get involved in the activities of the group but remain a passive observer, watching and listening to its activities and drawing conclusions

For example, you might want to study the functions carried out by nurses in a hospital. As an observer, you could watch, follow and record the activities as they are performed.

After making a number of observations, conclusions could be drawn about the functions nurses carry out in the hospital.

Demerits of observation method

The observer may lose the objectivity to the extent he/she participates emotionally; the problem of observation-control is not solved; and it may narrow-down the researcher's range of experience.

Controlled and uncontrolled observation

If the observation takes place in the natural setting, it may be termed as uncontrolled observation, but when observation takes place according to definite pre-arranged plans, involving experimental procedure, the same is then termed controlled observation.

Interview Method

- The interview method of collecting data involves presentation of oral-verbal stimuli and reply in terms of oral-verbal responses.

This method can be used through

a) Personal interviews b) Telephone interviews.

(a) *Personal interviews*: Personal interview method requires a person known as the interviewer asking questions in a face-to-face contact to the other person or persons.

This sort of interview may be in the form of direct personal investigation or it may be indirect oral investigation.

In the case of direct personal investigation the interviewer has to collect the information personally from the sources concerned which is suitable for intensive investigations.

In certain cases it may not be possible or worthwhile to contact directly the persons concerned or on account of the extensive scope of enquiry, the direct personal investigation technique may not be used.

In an indirect oral examination can be conducted under which the interviewer has to cross-examine other persons who are supposed to have knowledge about the problem under investigation and the information

Most of the commissions and committees appointed by government to carry on investigations make use of this method.

The method of collecting information through personal interviews is usually carried out in a structured way

Structured interviews: Involve the use of a set of predetermined questions and of highly standardised techniques of recording.

The interviewer in a structured interview follows a rigid procedure laid down, asking questions in a form and order prescribed.

Unstructured interviews are characterised by a flexibility of approach to questioning and do not follow a system of pre-determined questions and standardised techniques of recording information.

Focussed interview : Focus attention on the given experience of the respondent and its effects so that interviewer has the freedom to decide the manner and sequence in which the questions would be asked and has also the freedom to explore reasons and motives.

The main task of the interviewer in case of a focussed interview is to confine the respondent to a discussion of issues with which he seeks conversance.

These interviews are used in the development of hypotheses and constitute a major type of unstructured interviews.

- The *clinical interview* is concerned with broad underlying feelings or motivations or with the course of individual's life experience.
- The method of eliciting information under it is generally left to the interviewer's discretion.
- In case of *non-directive interview*, the interviewer's function is simply to encourage the respondent to talk about the given topic with a bare minimum of direct questioning.

Merits of Personal Interview

- (i) More information and that too in greater depth can be obtained.
- (ii) Interviewer by his own skill can overcome the resistance, if any, of the respondents; the interview method can be made to yield an almost perfect sample of the general population.
- (iii) There is greater flexibility under this method as the opportunity to restructure questions is always there, specially in case of unstructured interviews.

- (iv) Observation method can as well be applied to recording verbal answers to various questions.
- (v) Personal information can as well be obtained easily under this method.
- (vi) Samples can be controlled more effectively as there arises no difficulty of the missing returns; non-response generally remains very low.
- (vii) The interviewer can usually control which person(s) will answer the questions. This is not possible in mailed questionnaire approach. If so desired, group discussions may also be held

- viii) The interviewer may catch the informant off-guard and thus may secure the most spontaneous reactions than would be the case if mailed questionnaire is used.
- (ix) The language of the interview can be adopted to the ability or educational level of the person interviewed and as such misinterpretations concerning questions can be avoided.
- (x) The interviewer can collect supplementary information about the respondent's personal characteristics and environment which is often of great value in interpreting results.

Demerits of Personal Interviews

- (i) It is a very expensive method, specially when large and widely spread geographical sample is taken.
- (ii) There remains the possibility of the bias of interviewer as well as that of the respondent; there also remains the headache of supervision and control of interviewers.
- (iii) Certain types of respondents such as important officials or executives or people in high income groups may not be easily approachable under this method and to that extent the data may prove inadequate.

- (iv) This method is relatively more-time-consuming, specially when the sample is large and recalls upon the respondents are necessary.
- (v) The presence of the interviewer on the spot may over-stimulate the respondent, sometimes even to the extent that he may give imaginary information just to make the interview interesting.
- (vi) Under the interview method the organisation required for selecting, training and supervising the field-staff is more complex with formidable problems.

(vii) Interviewing at times may also introduce systematic errors.

(viii) Effective interview presupposes proper rapport with respondents that would facilitate free and frank responses. This is often a very difficult requirement.

Telephone Interviews

This method of collecting information consists in contacting respondents on telephone itself and plays important part in industrial surveys, in developed regions.

Merits of Telephone Interviews

1. It is more flexible in comparison to mailing method.
2. It is faster than other methods i.e., a quick way of obtaining information.
3. It is cheaper than personal interviewing method; here the cost per response is relatively low.
4. Recall is easy; call backs are simple and economical.

5. There is a higher rate of response than what we have in mailing method; the non-response is generally very low.
6. Replies can be recorded without causing embarrassment to respondents.
7. Interviewer can explain requirements more easily.
8. At times, access can be gained to respondents who otherwise cannot be contacted for one reason or the other.
9. No field staff is required.
10. Representative and wider distribution of sample is possible.

Demerits of Telephone Interviews

1. Little time is given to respondents for considered answers; interview period is not likely to exceed five minutes in most cases.
2. Surveys are restricted to respondents who have telephone facilities.
3. Extensive geographical coverage may get restricted by cost considerations.
4. It is not suitable for intensive surveys where comprehensive answers are required to various questions.
5. Possibility of the bias of the interviewer is relatively more.
6. Questions have to be short and to the point; probes are difficult to handle.

COLLECTION OF DATA THROUGH QUESTIONNAIRES

- Data collection through Questionnaire is used in case of big enquiries and most extensively employed in various economic and business surveys.
- This method is being adopted by private individuals, research workers, private and public organisations and even by governments
- In this method a questionnaire is sent (usually by post) to the persons concerned with a request to answer the questions and return the questionnaire.
- A questionnaire consists of a number of questions printed or typed in a definite order on a form or set of forms.

Merits of Questionnaires method

1. There is low cost even when the universe is large and is widely spread geographically.
2. It is free from the bias of the interviewer; answers are in respondents' own words.
3. Respondents have adequate time to give well thought out answers.
4. Respondents, who are not easily approachable, can also be reached conveniently.
- 5 Large samples can be made use of and thus the results can be made more dependable and reliable.

Demerits of Questionnaires method

1. Low rate of return of the duly filled in questionnaires; bias due to no-response is often indeterminate.
2. It can be used only when respondents are educated and cooperating.
3. The control over questionnaire may be lost once it is sent.
4. There is inbuilt inflexibility because of the difficulty of amending the approach once questionnaires have been despatched.

5. There is also the possibility of ambiguous replies or omission of replies altogether to certain questions; interpretation of omissions is difficult.
6. It is difficult to know whether willing respondents are truly representative.
7. This method is likely to be the slowest of all.

- Before using this method, it is always advisable to conduct 'pilot study' (Pilot Survey) for testing the questionnaires. In a big enquiry the significance of pilot survey is felt very much
- Pilot survey is in fact the replica and rehearsal of the main survey.
- Pilot survey, being conducted by experts, brings to the light the weaknesses (if any) of the questionnaires and also of the survey techniques. From the experience gained in this way, improvement can be effected.

Main aspects of a questionnaire:

Questionnaire is considered as the heart of a survey operation. Hence it should be very carefully constructed. Otherwise the survey is bound to fail.

The main aspects of a questionnaire

- i) General form
- ii) Question sequence
- iii) Question formulation and wording.

i)General form: It can either be structured questionnaire or unstructured questionnaire.

Structured questionnaires are definite, concrete and pre-determined questions. The questions are presented with exactly the same wording and in the same order to all respondents.

Resort is taken to this sort of standardisation to ensure that all respondents reply to the same set of questions.

The form of the question may be either closed (i.e., of the type 'yes' or 'no') or open (i.e., inviting free response) but should be stated in advance and not constructed during questioning

Structured questionnaires may also have fixed alternative questions in which responses of the informants are limited to the stated alternatives.

Highly structured questionnaire is one in which all questions and answers are specified and comments in the respondent's own words are held to the minimum

When these characteristics are not present in a questionnaire, it can be termed as unstructured or non-structured questionnaire

In an unstructured questionnaire, the interviewer is provided with a general guide on the type of information to be obtained, but the exact question formulation is largely his own responsibility and the replies are to be taken down in the respondent's own words to the extent possible; in some situations recorders may be used to achieve this goal.

Structured questionnaires are simple to administer and relatively inexpensive to analyse.

Question sequence:

In order to make the questionnaire effective and to ensure quality to the replies received, a researcher should pay attention to the question-sequence in preparing the questionnaire.

A proper sequence of questions reduces considerably the chances of individual questions being misunderstood.

The question-sequence must be clear and smoothly-moving. ie The relation of one question to another should be readily apparent to the respondent, with questions that are easiest to answer being put in the beginning.

The first few questions are particularly important since they are likely to influence the attitude of the respondent and in seeking his desired cooperation.

The opening questions should be such as to arouse human interest.

Following the opening questions, we should have questions that are really vital to the research problem and a connecting thread should run through successive questions.

Ideally, the question sequence should conform to the respondent's way of thinking.

The following type of questions should be avoided as opening questions in a questionnaire:

1. Questions that put too great a strain on the memory or intellect of the respondent;
2. Questions of a personal character;
3. Questions related to personal wealth, etc.

Relatively difficult questions must be relegated towards the end so that even if the respondent decides not to answer such questions, considerable information would have already been obtained.

Question-sequence should go from the general to the more specific and the researcher must remember that the answer to a given question is a function not only of the question itself, but of all previous questions as well.

For instance, if one question deals with the price paid for coffee and the next with reason for preferring that particular brand, the answer to this latter question may be couched largely in terms of price differences

Question formulation and wording:

Each question must be very clear for any sort of misunderstanding can do irreparable harm to a survey.

Question should be impartial in order not to give a biased picture of the true state of affairs.

Questions should be constructed with a view to their forming a logical part of a well thought out tabulation plan.

In general, all questions should meet the following standards

- (a) should be easily understood;
- (b) should be simple i.e., should convey only one thought at a time;
- (c) should be concrete and should conform as much as possible to the respondent's way of thinking.

For instance, instead of asking. "How many razor blades do you use annually?" The more realistic question would be to ask, "How many razor blades did you use last week?"

There are two principal forms of questions:

- i) Multiple choice question

- ii) The open-end question.

In the multiple choice question the respondent selects one of the alternative possible answers put to him.

In the open-end question he/she has to supply the answer in his own words.

The question with only two possible answers (usually 'Yes' or 'No') can be taken as a special case of the multiple choice question, or can be named as a 'closed question.'

Advantages of Multiple choice or Closed form questions

- i) Easy handling
- ii) Simple to answer
- iii) Quick and relatively inexpensive to analyse.
- iv) Most amenable to statistical analysis.
- v) Sometimes, the provision of alternative replies helps to make clear the meaning of the question.

The main drawback of fixed alternative questions is
“putting answers in people’s mouths”

i.e., they may force a statement of opinion on an issue
about which the respondent does not in fact have
any opinion.

They are not appropriate when the issue under
consideration happens to be a complex one and also
when the interest of the researcher is in the
exploration of a process.

In an **open-ended question** the possible responses are *not* given. In the case of a questionnaire, the respondent writes down the answers in his/her words, but in the case of an interview schedule the investigator records the answers either verbatim or in a summary.

In a **closed question** the possible answers are set out in the questionnaire or schedule and the respondent or the investigator ticks the category that best describes the respondent's answer. It is usually wise to provide a category 'Other/please explain' to accommodate any response not listed. The questions in Figure 1 are classified as closed questions. The same questions could be asked as open-ended questions, as shown in Figure 2.

A. Please indicate your age by placing a tick in the appropriate category.

- Under 15 ☐
- 15–19 years ☐
- 20–24 years ☐

B. How would you describe your current marital status?

- Married ☐
- Single ☐
- De facto ☐
- Divorced ☐
- Separated ☐

C. What is your average annual income?

- Under \$10 000 ☐
- \$10 000–\$19 999 ☐
- \$20 000–\$29 999 ☐
- \$30 000–\$39 999 ☐
- \$40 000+ ☐

OR

C(a). How would you categorise your average annual income?

- Above average ☐
- Average ☐
- Below average ☐

D. What, in your opinion, are the qualities of a good administrator?

- Able to make decisions ☐
- Fast decision maker ☐
- Able to listen ☐
- Impartial ☐
- Skilled in interpersonal communication ☐
- Other, please specify

Examples of closed questions

A. Please indicate your age by placing a tick in the appropriate category.

Under 15 ☐

15–19 years ☐

20–24 years ☐

B. How would you describe your current marital status?

Married ☐

Single ☐

De facto ☐

Divorced ☐

Separated ☐

C. What is your average annual income?

Under \$10 000 ☐

\$10 000–\$19 999 ☐

\$20 000–\$29 999 ☐

\$30 000–\$39 999 ☐

\$40 000+ ☐

OR

C(a). How would you categorise your average annual income?

Above average ☐

Average ☐

Below average ☐

D. What, in your opinion, are the qualities of a good administrator?

Able to make decisions ☐

Fast decision maker ☐

Able to listen ☐

Impartial ☐

Skilled in interpersonal communication ☐

Other, please specify

- A. What is your current age? _____ years
- B. How would you describe your current marital status? _____
- C. What is your average annual income? \$ _____
- D. What, in your opinion, are the qualities of a good administrator?
- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____

Examples of open-ended questions

Let us consider example, the question about the variable:
‘income’.

In closed questions income can be qualitatively recorded in categories such as ‘above average/average/below average’, or quantitatively in categories such as ‘under \$10 000/\$10 000–\$19 999/...’.

The choice of qualitative and quantitative categories affects the unit of measurement for income (qualitative uses the ordinal scale and quantitative the ratio scale of measurement), which in turn will affect the application of statistical procedures.

For example, you cannot calculate the average income of a person from the responses to question C(a) in Figure ; nor can you calculate the median or modal category of income. But from the responses to question C, you can accurately calculate modal category of income

The average and the median income cannot be accurately calculated (such calculations are usually made under certain assumptions). From the responses to question C in Figure 2, where the income for a respondent is recorded in exact dollars, the different descriptors of income can be calculated very accurately.

In addition, information on income can be displayed in any form. You can calculate the average, median or mode. The same is true for any other information obtained in response to open-ended and closed questions.

COLLECTION OF DATA THROUGH SCHEDULES

This method of data collection is like the collection of data through questionnaire, with little difference which lies in the fact that schedules (proforma containing a set of questions) are being filled in by the enumerators who are specially appointed for the purpose.

These enumerators along with schedules, go to respondents, put to them the questions from the proforma in the order the questions are listed and record the replies in the space meant for the same in the proforma.

In certain situations, schedules may be handed over to respondents and enumerators may help them in recording their answers to various questions in the said schedules.

Enumerators explain the aims and objects of the investigation and also remove the difficulties which any respondent may feel in understanding the implications of a particular question or the definition or concept of difficult terms.

This method requires the selection of enumerators for filling up schedules or assisting respondents to fill up schedules and as such enumerators should be very carefully selected.

The enumerators should be trained to perform their job well and the nature and scope of the investigation should be explained to them thoroughly so that they may well understand the implications of different questions put in the schedule.

Enumerators should be intelligent and must possess the capacity of cross examination in order to find out the truth. Above all, they should be honest, sincere, hardworking and should have patience and perseverance.

This method of data collection is very useful in extensive enquiries and can lead to fairly reliable results.

It is very expensive and is usually adopted in investigations conducted by governmental agencies or by some big organisations. Population census all over the world is conducted through this method.

COLLECTION OF SECONDARY DATA

Secondary data refer to the data which have already been collected and analysed by someone else.

When the researcher utilises secondary data, then s/he has to look into various sources from where s/he can obtain them

Secondary data may either be published data or unpublished data.

Published data are available in:

- (a) Various publications of the central, state and local governments;
- (b) Various publications of foreign governments or of international bodies and their subsidiary organisations;
- (c) Technical and trade journals;
- (d) Books, magazines and newspapers;
- (e) Reports and publications of various associations connected with business and industry, banks, stock exchanges, etc.;
- (f) Reports prepared by research scholars, universities, economists, etc. in different fields;
- (g) Public records and statistics, historical documents, and other sources of published information.

Before using secondary data, the researcher must see that they possess following characteristics:

1. Reliability of data: The reliability can be tested by finding out such things about the said data:
 - (a) Who collected the data?
 - (b) What were the sources of data?
 - (c) Were they collected by using proper methods
 - (d) At what time were they collected?
 - (e) Was there any bias of the compiler?
 - (f) What level of accuracy was desired? Was it achieved

2. Suitability of data:

The data that are suitable for one enquiry may not necessarily be found suitable in another enquiry. Hence, if the available data are found to be unsuitable, they should not be used by the researcher.

The researcher must very carefully scrutinize the definition of various terms and units of collection used at the time of collecting the data from the primary source originally.

The object, scope and nature of the original enquiry must also be studied. If the researcher finds differences in these, the data will remain unsuitable for the present enquiry and should not be used.

3. Adequacy of data:

If the level of accuracy achieved in data is found inadequate for the purpose of the present enquiry, they will be considered as inadequate and should not be used by the researcher.

The data will also be considered inadequate, if they are related to an area which may be either narrower or wider than the area of the present enquiry

It is very risky to use the already available data. The already available data should be used by the researcher only when he finds them reliable, suitable and adequate

SELECTION OF APPROPRIATE METHOD FOR DATA COLLECTION

The researcher select the methods for his own study, based on the following factors

1. Nature, scope and object of enquiry:

The method selected should be suits the type of enquiry that is to be conducted by the researcher.

This factor is also important in deciding whether the data available (secondary data) are to be used or the data not yet available (primary data) are to be collected

2.Availability of funds:

Availability of funds for the research project determines to a large extent the method to be used for the collection of data.

When funds at the disposal of the researcher are very limited, s/he will have to select a comparatively cheaper method which may not be as efficient and effective as some other costly method.

Finance, is a big constraint in practice and the researcher has to act within this limitation.

3. Time factor:

Availability of time has to be taken into account in deciding a particular method of data collection.

Some methods take relatively more time, whereas with others the data can be collected in a comparatively shorter duration.

The time at the disposal of the researcher affects the selection of the method by which the data are to be collected.

4. Precision required:

Precision required is another important factor to be considered at the time of selecting the method of collection of data.

Each method of data collection has its uses and none is superior in all situations.

For instance, telephone interview method may be considered appropriate (assuming telephone population) if funds are restricted, time is also restricted and the data is to be collected in respect of few items with or without a certain degree of precision.

In case funds permit and more information is desired, personal interview method may be said to be relatively better.

In case time is ample, funds are limited and much information is to be gathered with no precision, then mail-questionnaire method can be regarded more reasonable.

When funds are ample, time is also ample and much information with no precision is to be collected, then either personal interview or the mail-questionnaire or the joint use of these two methods may be taken as an appropriate method of collecting data.

To cover a wide geographic area, the use of mail-questionnaires supplemented by personal interviews will yield more reliable results per rupee spent than either method alone.

The secondary data may be used in case the researcher finds them reliable, adequate and appropriate for his research.

While studying motivating influences in market researches or studying people's attitudes in psychological/social surveys, we can resort to the use of one or more of the projective techniques stated earlier.

Such techniques are of immense value in case the reason is obtainable from the respondent who knows the reason but does not want to admit it or the reason relates to some underlying psychological attitude and the respondent is not aware of it.

The most desirable approach with regard to the selection of the method depends on the nature of the particular problem and on the time and resources (money and personnel) available along with the desired degree of accuracy.

Above all this, much depends upon the ability and experience of the researcher.

Dr. A.L. Bowley's remark in this context is very appropriate when he says that "in collection of statistical data common sense is the chief requisite and experience the chief teacher."

Important data available for scientists in World Wide Web

Websites to find free, interesting datasets

- 1)FiveThirtyEight.
- 2)BuzzFeed News.
- 3)Kaggle.**
- 4)Socrata.
- 5)Awesome-Public-**Datasets** on Github.
- 6)Google Public **Datasets**.
- 7)UCI Machine Learning Repository.
- 8)Data.gov
- 9)Academic Torrents
- 10)Quandl
- 11)Jeremy Singer-vine

If you're new to the data space, or if you've recently learned a new skill, or just trying to build a more robust data science/analyst portfolio, a perfect way of solidifying your skills is to do some mini-projects focused on your new skills.

Below we outline a few places you can find publicly available data for your next project.

If you're interested in practicing real data scientist and analyst interview questions, feel free to sign up for our email newsletter, where we send a few curated questions per week to help you prepare for interviews at top companies

1. FiveThirtyEight

FiveThirtyEight is an interactive news and sports site that has some incredible data visualizations ([which you should totally check out](#)). They makes a lot of their data open to the public, meaning you can download and play with the source data yourself!

Here are some examples:

- [Airline Safety](#) — contains information on accidents from each airline
- [US Weather History](#) — historical weather data for the US.
- [Study Drugs](#) — data on who's taking Adderall in the US.

2. BuzzFeed News

BuzzFeed makes the data sets, analysis, libraries, tools, and guides used in its articles available on Github. Check them out to learn from some of the best!

Here are some examples:

- [Federal Surveillance Planes](#) — contains data on planes used for domestic surveillance.
- [Zika Virus](#) — data about the geography of the Zika virus outbreak.
- [Firearm background checks](#) — data on background checks of people attempting to buy firearms.

3. Kaggle

Kaggle, [recently acquired by Google](#), is a place where you can learn, practice, and fine-tune your data science/analytics skills. They have tons of data that's open to the public, and allow users of the platform to share code so you can learn best practices within the data space.

They also host [competitions](#) where you can win real money if you have a top ranking model!

Here are some examples:

- [Federal Surveillance Planes](#) — contains data on planes used for domestic surveillance.
- [Zika Virus](#) — data about the geography of the Zika virus outbreak.
- [Firearm background checks](#) — data on background checks of people attempting to buy firearms.

4. Socrata

Socrata hosts cleaned open source data sources ranging from government, business, and education data sets.

Here are some examples:

- [White House staff salaries](#) — data on what each White House staffer made in 2010.
- [Radiation Analysis](#) — data on what milk products in what locations in the US were radioactive.
- [Workplace fatalities by US state](#) — the number of workplace deaths across the US.

5. Awesome-Public-Datasets on Github

This github hosts a library of awesome, public datasets! They are all sorted by category and link you straight to the hosting website.

Here are some examples:

- [Global Climate Data](#) — climate information for every country in the world with historical data in some cases date back to 1929
- [Heart rate time series data](#) — two series of data contains 1800 evenly-spaced measurements of instantaneous heart rate from a single subject
- [Plane crash database](#) — plane crash data dating from 1929 to now.

6. Google Public Datasets

Google lists all of the data sets on a page. Google has a cloud hosting service called Google Cloud Platform (GCP), and you can query using a tool called BigQuery to explore these datasets. You'll need to sign up for a GCP account, but the first 1TB of queries you make are free! But be careful not to go over or you'll have to pay!

Here are some examples:

- [US name data set](#) — contains all names from social security card applications from births that occur after 1879
- [Major League Baseball data](#) — data includes pitch-by-pitch data for Major League Baseball (MLB) games in 2016

7. UCI Machine Learning Repository

University of California Irvine hosts 440 data set as a service to the machine learning community. These data sets are nice because most of them are squeaky clean, and are ready for modeling!

Here are some examples:

- [Iris data set](#) — the most famous pattern recognition dataset.
- [Wine data set](#) — using chemical analysis to determine the origin of wine.
- [Forest fires](#) — try to predict the burn area of forest fires using this dataset.

8. [Data.gov](#)

Data.gov allows you to download and explore data from multiple US government agencies.

Data can range from government budgets to climate data. The data is very well documented so you should have an easy time to navigate the sources.

You can browse the data sets on Data.gov directly, without registering. You can browse by topic area, or search for a specific data set.

Here are some examples:

- [Food Environment Atlas](#) — contains data on how local food choices affect diet in the
- [School system finances](#) — a survey of the finances of school systems in the US.
- [Chronic disease data](#) — data on chronic disease indicators in areas across the US.

9. Academic Torrents

Academic Torrents is a site that is geared around sharing the data sets from scientific papers. It has tons of interesting data sets. You can browse the data sets directly on the site, and download if you find interesting!

Here are some examples:

- [Enron emails](#) — a set of many emails from executives at Enron, a company that famously went bankrupt.
- [Student learning factors](#) — a set of factors that measure and influence student learning.
- [News articles](#) — contains news article attributes and a target variable.

10. Quandl

Quandl is a repository of economic and financial data. Some of the datasets are free, while others are up for purchase.

Here are some examples:

- [Entrepreneurial activity by race and other factors](#) — contains data from the Kauffman foundation on entrepreneurs in the US.
- [Chinese macroeconomic data](#) — indicators of Chinese economic health.
- [US Federal Reserve data](#) — US economic indicators, from the Federal Reserve.

11. Jeremy Singer-Vine

Jeremy Singer-Vine collects awesome data sets across multiple sources. If you're interested in getting data sets straight to your inbox, you should consider signing up for his [newsletter](#).

Interested in practicing for data scientist or analyst interviews?

We send 3 questions each week to thousands of data scientists and analysts preparing for interviews or just keeping their skills sharp. You can sign up to receive the questions for free on our [home page](#).

RES7001- RM - DATA ANALYSIS

Module – 5

Dr.K.Karthikeyan

Professor, Dept. of Maths,

SAS, VIT Vellore

Classification of data

- The collected data, also known as raw data or ungrouped data are always in an unorganized form and need to be organised and presented in meaningful and readily comprehensible form in order to facilitate further statistical analysis.
- It is essential for an investigator to condense a mass of data into more and more comprehensible and assimilable form.
- The process of grouping into different classes or sub classes according to some characteristics is known as classification, tabulation is concerned with the systematic arrangement and presentation of classified data.
- Classification is the first step in tabulation.
- For Example, letters in the post office are classified according to their destinations viz., Delhi, Madurai, Bangalore, Mumbai etc.,

Principal Objects of Classification data

- To condense the mass of data in such a manner that similarities and dissimilarities can be readily apprehended. Millions of figures can thus be arranged in a few classes having common features.
- To facilitate comparison
- To pinpoint the most significant features of the data at a glance
- To give prominence to the important information gathered while dropping out the unnecessary elements
- To enable a statistical treatment of the material collected

Objects of Classification:

The following are main objectives of classifying the data:

1. It condenses the mass of data in an easily assimilable form.
2. It eliminates unnecessary details.
3. It facilitates comparison and highlights the significant aspect of data.
4. It enables one to get a mental picture of the information and helps in drawing inferences.
5. It helps in the statistical treatment of the information collected.

Types of Classification

Broadly, the data can be classified on the basis of following four criteria:

- Geographical, *i.e.*, area-wise, *e.g.*, cities, districts, etc.
- Chronological, *i.e.*, on the basis of time.
- Qualitative, *i.e.*, according to some attributes.
- Quantitative, *i.e.*, in terms of magnitudes.

Geographical Classification

In this type of classification data are classified on the basis of geographical or vocational differences between the various items, like Countries, States, cities, regions, zones, areas, etc. For instance, the data about the production and per capita availability of milk in India for the years 2005-06 to 2011-12 is given in the following table :

PRODUCTION AND PER CAPITA AVAILABILITY OF MILK

Year	Per Capita Availability (Grams/day)	Production (million tonnes)
2005-06	241	97.1
2006-07	251	102.6
2007-08	260	107.9
2008-09	266	112.2
2009-10	273	116.4
2010-11	281	121.8
2011-12	290	127.9

Geographical classification is usually listed in alphabetical order for easy reference. Items may also be listed by size to emphasise the important areas as in ranking the States by population. Normally, in reference table the first approach is followed and in summary tables the second approach is followed.

In this type of classification the data are classified according to geographical region or place. For instance, the production of paddy in different states in India, production of wheat in different countries etc.,

Country	America	China	Denmark	France	India
Yield of wheat in (kg/acre)	1925	893	225	439	862

Chronological Classification

When data are observed over a period of time the type of classification is known as chronological classification. For example, we may present the figures of population (or production, sales, etc.) as follows:

POPULATION OF INDIA FROM 1951 TO 2011

Year	Population (in crore)	Year	Population (in crore)
1951	36.11	1991	84.64
1961	43.92	2001	102.87
1971	54.82	2011	121.00
1981	68.33		

Time series are usually listed in chronological order, normally starting with the earliest period. When the major emphasis falls on the most recent events, a reverse time order may be used.

In chronological classification the collected data are arranged according to the order of time expressed in years, months, weeks, etc., The data is generally classified in ascending order of time. For example, the data related with population, sales of a firm, imports and exports of a country are always subjected to chronological classification.

The estimates of birth rates in India during 1970 – 76 are

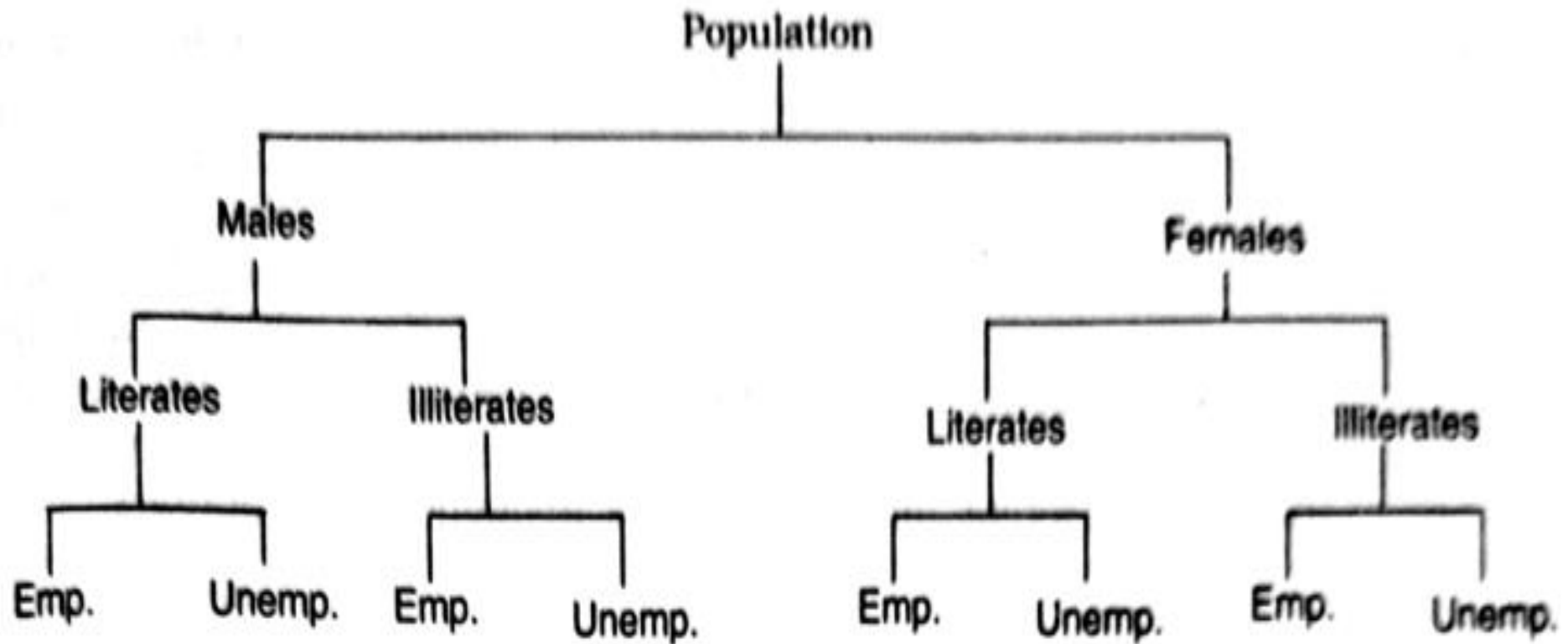
Year	1970	1971	1972	1973	1974	1975	1976
Birth Rate	36.8	36.9	36.6	34.6	34.5	35.2	34.2

Qualitative Classification

In qualitative classification, data are classified on the basis some attribute or quality such as sex, colour of hair, literacy, religion, etc.,

In this type the attribute under study cannot be measured but can find out whether it is present or absent in the units of the production under study

For example the population under study be divided as manifold classification



Note. Emp. indicates Employed and Unemp. indicates Unemployed.

Quantitative Classification

Quantitative classification refers to the classification of data according to some characteristics that can be measured, such as height, weight, income, sales, profits, production, etc. For example, the students of a college may be classified according to weight as follows:

<i>Weight (in lbs.)</i>	<i>No. of Students</i>
90-100	50
100-110	200
110-120	260
120-130	360
130-140	90
140-150	40
Total	1,000

Such a distribution is known as empirical frequency distribution or simple frequency distribution.

In this type of classification, there are two elements, namely (i) the variable, *i.e.*, the weight in the above example, and (ii) the frequency, *i.e.*, the number of students in each class. There were 50 students having weight ranging from 90 to 100 lbs, 200 students having weight ranging from 100 to 110 lbs, and so on. Thus we can find out the ways in which the frequencies are distributed.

Tabulation

Tabulation is the process of summarizing classified or grouped data in the form of a table so that it is easily understood and an investigator is quickly able to locate the desired information.

A table is a systematic arrangement of classified data in columns and rows.

A statistical table makes it possible for the investigator to present a huge mass of data in a detailed and orderly form.

It facilitates comparison and reveals certain patterns in data which are otherwise not obvious. Classification and 'Tabulation' are not two distinct processes. They go together.

Before tabulation data are classified and then displayed under different columns and rows of a table

Advantages of Tabulation

Statistical data arranged in a tabular form serve following objectives:

1. It simplifies complex data and the data presented are easily understood.
2. It facilitates comparison of related facts.
3. It facilitates computation of various statistical measures like averages, dispersion, correlation etc.
4. It presents facts in minimum possible space and unnecessary repetitions and explanations are avoided. Moreover, the needed information can be easily located.
5. Tabulated data are good for references and they make it easier to present the information in the form of graphs and diagrams

An ideal table should consist of the following main parts:

1. Table number
2. Title of the table
3. Captions or column headings
4. Stubs or row designation
5. Body of the table
6. Footnotes
7. Sources of data

Type of Tables:

Tables can be classified according to their purpose, stage of enquiry, nature of data or number of characteristics used. On the basis of the number of characteristics, tables may be classified as follows:

1. Simple or one-way table
2. Two way table
3. Manifold table

Simple or one-way Table:

A simple or one-way table is the simplest table which contains data of one characteristic only. A simple table is easy to construct and simple to follow. For example, the blank table given below may be used to show the number of adults in different occupations in a locality.

The number of adults in different occupations in a locality

Occupations	No. of Adults
Total	

Two-way Table:

A table, which contains data on two characteristics, is called a twoway table. In such case, therefore, either stub or caption is divided into two co-ordinate parts. In the given table, as an example the caption may be further divided in respect of ‘ sex’ . This subdivision is shown in two-way table, which now contains two characteristics namely, occupation and sex.

The number of adults in a locality in respect of occupation and sex

Occupation	No. of Adults		Total
	Male	Female	
Total			

Manifold Table:

Thus, more and more complex tables can be formed by including other characteristics. For example, we may further classify the caption sub-headings in the above table in respect of “marital status”, “ religion” and “socio-economic status” etc. A table ,which has more than two characteristics of data is considered as a manifold table. For instance , table shown below shows three characteristics namely, occupation, sex and marital status.

Occupation	No. of Adults						Total
	Male			Female			
	M	U	Total	M	U	Total	
Total							

Foot note: M Stands for Married and U stands for unmarried.

Manifold tables, though complex are good in practice as these enable full information to be incorporated and facilitate analysis of all related facts. Still, as a normal practice, not more than four characteristics should be represented in one table to avoid confusion. Other related tables may be formed to show the remaining characteristics

2.20 FREQUENCY DISTRIBUTION

Frequency Distribution. *Frequency distribution of a variable x is the ordered set $\{x, f\}$, where f is the frequency.* It shows all scores in a set of data, together with the frequency of each class. When the data is presented in frequency distribution form, one can easily understand the information contained in the raw data.

Frequency distribution are of two types :

- (i) Discrete Frequency Distribution**
- (ii) Grouped (or Continuous) Frequency Distribution**

2.20.1 Discrete Frequency Distribution

The construction of discrete frequency distribution from the given raw data is done by the method of tally marks as explained earlier.

2.20.2 Construction of Discrete Frequency Distribution Table

The frequency distribution table has three columns headed by

- | | | |
|---------------------------|------------------------|---------------|
| 1. Variables (or Classes) | 2. Tally Marks or Bars | 3. Frequency. |
|---------------------------|------------------------|---------------|

The table is constructed by the following steps :

- Step 1.** Prepare three columns, viz., one for the **variable** (or classes), another for **tally marks** and the **third** for the **frequency** corresponding the variable (or class).
- Step 2.** Arrange the given data (or values) from the lowest to the highest in the first column under the heading **variable** (or classes).
- Step 3.** Take the first observation in the raw data and put a bar (or vertical line |) in the second column under **Tally Marks** opposite to it. Then take a second observation and put a **tally mark** opposite to it. Continue this process till all the observations of the given raw data are exhausted. For the sake of convenience, record the tally marks in **bunches of five**, the **fifth bar is placed diagonally crossing the other four** (5 is represented by $\overline{||||}$) leave some space between each block of bars.
- Step 4.** Count the **tally marks** of column 2 and place this number opposite to the value of the variable in the third column headed by **Frequency**.
- Step 5.** Give a suitable title to the frequency distribution table so that it exactly conveys the information contained in the table.

Example 14. Form a discrete frequency distribution from the following scores :

15, 18, 16, 20, 25, 24, 25, 20, 16, 15, 18, 18, 16, 24, 15, 20, 28, 30, 27, 16, 24, 25, 20, 18, 28, 27, 25, 24, 24, 18, 18, 25, 20, 16, 15, 20, 27, 28, 29, 16.

Solution.

Table : Frequency Distribution of Scores

<i>Variate</i>	<i>Tally Marks</i>	<i>Frequency</i>
15	IIII	4
16	IIII I	6
18	IIII I	6
20	IIII I	6
24	IIII	5
25	IIII	5
27	III	3
28	III	3
29	I	1
30	I	1
Total	Dr.K.Karthikeyan SAS	40

In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.

1	0	3	2	1	5	6	2
2	1	0	3	4	2	1	6
3	2	1	5	3	3	2	4
2	2	3	0	2	1	4	5
3	3	4	4	1	2	4	5

Represent the data in the form of a discrete frequency distribution.

Solution:

Frequency distribution of the number of children

Number of Children	Tally Marks	Frequency
0		3
1	 	7
2	 	10
3	 	8
4	 	6
5		4
6		2
	Total	40

Types of class intervals

There are three methods of classifying the data according to class intervals namely

- a) Exclusive method
- b) Inclusive method
- c) Open-end classes

a) Exclusive method:

When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class; it is known as the exclusive method of classification.

Expenditure (Rs.)	No. of families
0-5000	60
5000-10000	95
10000-15000	122
15000-20000	83
20000-25000	40
Total	400

It is clear that the exclusive method ensures continuity of data as much as the upper limit of one class is the lower limit of the next class. In the above example, there are 60 families whose expenditure is between Rs.0 and Rs.4999.99. A family whose expenditure is Rs.5000 would be included in the class interval 5000-10000. This method is widely used in practice.

b) Inclusive method:

In this method, the overlapping of the class intervals is avoided. Both the lower and upper limits are included in the class interval. This type of classification may be used for a grouped frequency distribution for discrete variable like members in a family, number of workers in a factory etc., where the variable may take only integral values. It cannot be used with fractional values like age, height, weight etc.

This method may be illustrated as follows:

Class interval	Frequency
5-9	7
10-14	12
15-19	15
20-29	21
30-34	10
35-39	5
Total	70

Thus to decide whether to use the inclusive method or the exclusive method, it is important to determine whether the variable under observation is a continuous or discrete one. In case of continuous variables, the exclusive method must be used. The inclusive method should be used in case of discrete variable.

c) Open end classes:

A class limit is missing either at the lower end of the first class interval or at the upper end of the last class interval or both are not specified. The necessity of open end classes arises in a number of practical situations, particularly relating to economic and medical data when there are few very high values or few very low values which are far apart from the majority of observations.

The example for the open-end classes as follows :

Salary Range	No. of workers
Below 2000	7
2000-4000	5
4000-6000	6
6000-8000	4
8000 and above	3

Thus the number of class intervals can be fixed arbitrarily keeping in view the nature of problem under study or it can be decided with the help of Sturges' Rule. According to him, the number of classes can be determined by the formula

$$K = 1 + 3.322 \log_{10} N$$

Where N = Total number of observations

log = logarithm of the number

K = Number of class intervals.

Size of class interval

Since the size of the class interval is inversely proportional to the number of class interval in a given distribution. The approximate value of the size (or width or magnitude) of the class interval 'C' is obtained by using sturges rule as

$$\begin{aligned}\text{Size of class interval} = C &= \frac{\text{Range}}{\text{Number of class interval}} \\ &= \frac{\text{Range}}{1 + 3.322 \log_{10} N}\end{aligned}$$

Where Range = Largest Value – smallest value in the distribution.

2.22-1 Construction of a Continuous Frequency Distribution Table

- Step 1.** Find the maximum and minimum value of the variate occurring in the data.
- Step 2.** Decide the number of classes to be formed. Note that the number of classes should be in range of 5 to 15.
- Step 3.** Find the difference between the maximum value and minimum value, i.e., find range. Divide this difference by the number of classes (as per **STEP II**) in order to determine the class interval.
- Step 4.** Take each item from the data, one at a time and put a tally mark (I) under the column Tally Mark against the class to which the item belong. Exhaust all the items of the given data in this form.
- Step 5.** Count the tally marks of column 2 headed by 'frequency' and place this number against the class to which the item belongs.
- Step 6.** Give suitable title to the frequency distribution table so that it conveys exactly the information contained in it.

Illustration 4. Prepare a frequency distribution of the marks obtained out of 100 for the following observations:

15	45	40	42	50	60	62	68	70	42
75	75	80	81	25	26	31	32	78	45
31	45	42	43	55	56	78	80	81	62
60	62	58	69	70	45	50	56	72	58
75	62	62	65	60	70	35	37	40	55

(B.Com., Bharthidasan Univ., 2009)

Solution. Since the lowest value is 15 and the largest 81, we take class interval of 10.

FREQUENCY DISTRIBUTION OF THE MARKS

<i>Marks</i>	<i>Tally Bars</i>	<i>Frequency</i>
15–25		1
25–35		5
35–45		8
45–55		6
55–65		14
65–75		7
75–85		9
Dr.K.Karthikeyan SAS		Total 50

Illustration 6. The marks obtained by 50 students are given below:

31	13	46	31	30	45	38	42	30	9
30	30	46	36	2	41	44	18	29	63
44	30	19	5	44	15	7	25	12	30
6	22	24	37	15	6	39	32	21	20
42	31	19	14	23	28	17	53	22	21

Construct a group frequency distribution.

(M.Com., Calicut Univ., 2009)

Solution. The lowest value is 2 and largest 63. The appropriate class intervals shall be 10 because 7 classes would be formed by taking 10 as class interval.

FREQUENCY DISTRIBUTION OF MARKS

Marks	Tally Bars	No. of Students
0-10		6
10-20		9
20-30		10
30-40		14
40-50		9
50-60		1
60-70		1
Total		50

Example 16. The water-tax bills (in rupees) of 30 houses in a locality are given below :

144, 184, 130, 195, 132, 134, 196, 114, 212, 174, 188, 210, 202, 145, 175, 154, 174, 178, 166, 146, 135, 115, 120, 114, 140, 188, 176, 166, 210, 208.

Construct an exclusive frequency distribution table with class size 10.

Solution. Minimum observation = 114, Maximum observation = 212.

\therefore Range = (maximum – minimum) = (212 – 114) = 98.

Class size = 10

Numbers of class intervals = $\frac{\text{Range}}{\text{Class size}} = \frac{212-114}{10} = \frac{98}{10} = 9.8$ or roughly 10.

The classes of equal size, covering the above data are :

114 – 124, 124 – 134, 134 – 144, 144 – 154, 154 – 164, 164 – 174, 174 – 184, 184 – 194, 194 – 204 and 204 – 214.

The frequency distribution table may be presented as shown below :

Table : Frequency distribution of Water Tax Bills

<i>Bills (in Rs.)</i>	<i>Tally Marks</i>	<i>Frequency</i>
114 – 124		4
124 – 134		2
134 – 144		3
144 – 154		3
154 – 164		1
164 – 174		2
174 – 184		5
184 – 194		3
194 – 204		3
204 – 214		4
Total	Dr.K.Karthikeyan SAS	30

Example 18. From the following observations prepare a classified frequency distribution by exclusive method.

110 108 126 132 149 136 125 112 138 155 125
 138 136 130 120 148 140 125 119 111 154 147
 165 137 145 132 150 137 142 135 125 126.

Solution. The lowest value is 108 and the highest value 165. The difference between the two extreme is 57. If we take class interval of 10, there will be six classes, starting from 100 – 110, 110 – 120 and so on.

Table : Frequency Distribution

<i>Class intervals</i>	<i>Tally Bars</i>	<i>Frequency</i>
100 – 110	I	1
110 – 120	IIII	4
120 – 130	IIII II	7
130 – 140	IIII III	10
140 – 150	IIII I	6
150 – 160	III	3
160 – 170	I	1
		N = 32

Example 19. Form a frequency distribution from the following data by inclusive method taking 4 as the magnitude of class intervals :

10	17	15	22	11	16	19	24	29	18
25	26	32	14	17	20	23	27	30	12
15	18	24	36	18	15	21	28	33	38
34	13	10	16	20	22	29	19	23	31.

Solution.

Table : Frequency Distribution

<i>Class intervals</i>	<i>Tally Bars</i>	<i>Frequency</i>
10 – 13	■	5
14 – 17	■ ■ ■ ■ ■	8
18 – 21	■ ■ ■ ■ ■	8
22 – 25	■ ■ ■ ■ ■	7
26 – 29	■	5
30 – 33	■	4
34 – 37	■	2
38 – 41	■	1
Total		40

Note : Lower limit of the first class is 10 as a minimum value of the variable is 10. The magnitude of the class interval is given to be 4. Under the inclusive method the various class intervals are 10 – 13, 14 – 17, 18 – 21, ..., 30 – 33, 34 – 37 and 38 – 41. To form a frequency distribution, first take the value 10, and put a tally mark against the class 10 – 13, for the second value 17, put a tally mark against the class 14 – 17, for the value 15, put a tally mark against the class 14 – 17, and for the value 22, put a tally mark against the class 22 – 25 and so on.

Given below are the number of tools produced by workers in a factory.

43	18	25	18	39	44	19	20	20	26
40	45	38	25	13	14	27	41	42	17
34	31	32	27	33	37	25	26	32	25
33	34	35	46	29	34	31	34	35	24
28	30	41	32	29	28	30	31	30	34
31	35	36	29	26	32	36	35	36	37
32	23	22	29	33	37	33	27	24	36
23	42	29	37	29	23	44	41	45	39
21	21	42	22	28	22	15	16	17	28
22	29	35	31	27	40	23	32	40	37

Construct frequency distribution with inclusive type of class interval. Also find.

1. How many workers produced more than 38 tools?
2. How many workers produced less than 23 tools?

Using sturges formula for determining the number of class intervals, we have

$$\begin{aligned}\text{Number of class intervals} &= 1 + 3.322 \log_{10} N \\ &= 1 + 3.322 \log_{10} 100 \\ &= 7.6\end{aligned}$$

$$\begin{aligned}\text{Size of class interval} &= \frac{\text{Range}}{\text{Number of class interval}} \\ &= \frac{46 - 13}{7.6} \\ &\approx 5\end{aligned}$$

Hence taking the magnitude of class intervals as 5, we have 7 classes 13-17, 18-22... 43-47 are the classes by inclusive type. Using tally marks, the required frequency distribution is obtained in the following table

Class Interval	Tally Marks	Number of tools produced (Frequency)
13-17		6
18-22		11
23-27		18
28-32		25
33-37		22
38-42		11
43-47		7
Total		100

Relative Frequency Distribution

At times it may be desirable to convert class frequencies to relative class frequencies to show the percentage of the total number of observations in each class.

In order to convert a frequency distribution to a relative frequency distribution, each of the class frequencies is divided by the total number of frequencies so that the relative frequencies would always total 1.

For Illustration 3 given above, the relative frequency distribution would be as follows:

<i>Marks</i>	<i>Frequency f</i>	<i>Relative Frequency</i>	<i>Obtained by Dividing Freq. by 50</i>
0-10	2	0.04	2 ÷ 50
10-20	5	0.10	5 ÷ 50
20-30	4	0.08	4 ÷ 50
30-40	5	0.10	etc.
40-50	8	0.16	
50-60	8	0.16	
60-70	7	0.14	
70-80	5	0.10	
80-90	4	0.08	
90-100	2	0.04	

N = 50

Dr.K.Karthikeyan SAS

Bivariate or Two way Frequency Distribution

Frequency distributions involving only one variable called as univariate frequency distributions.

In many real time problems, simultaneous study of two variables becomes necessary.

For example, classification of data relating to age of males and age of females, data relating to marks in statistics and marks in mathematics or heights and weights of students. The classified data based on two variables called bivariate frequency distribution.

In bivariate frequency distribution (X,Y) , where X is grouped in to m classes and Y is grouped in to n classes then the bivariate table will consists of mn cells.

Illustration 7. The data given below relate to the height and weight of 20 persons. You are required to form a two-way frequency table with class interval 62" to 64", 64" to 66", and so on and 115 to 125 lbs. 125 to 135 lbs., etc.

<i>S.No.</i>	<i>Weight</i>	<i>Height</i>	<i>S.No.</i>	<i>Weight</i>	<i>Height</i>
1	170	70	11	163	70
2	135	65	12	139	67
3	136	65	13	122	63
4	137	64	14	134	68
5	148	69	15	140	67
6	121	63	16	132	69
7	117	65	17	120	65
8	128	70	18	148	68
9	143	71	19	129	67
10	129	62	20	152	67

Solution. As per the requirements of the question, the population is to be divided into five classes according to the height of the persons included in each group and six classes according to the weight. There will thus be $5 \times 6 = 30$ cells.

For tabulating the information in appropriate cells, first, the row to which the height measurement (say, X), should belong is determined. Afterwards on a consideration of the weight (say, Y), the column in which it should be included is determined. The tabulation is recorded by tally bars. Thus the two-way table shall be prepared like this:”

**TWO-WAY FREQUENCY TABLE SHOWING WEIGHT AND
HEIGHT OF 20 PERSONS**

Weight in lbs (Y) Height in inches (X)	115–125	125–135	135–145	145–155	155–165	165–175	Total
62–64	(2)	(1)	—	—	—	—	3
64–66	(2)	—	(3)	—	—	—	5
66–68	—	(1)	(2)	(1)	—	—	4
68–70	—	(2)	—	(2)	—	—	4
70–72	—	(1)	(1)	—	(1)	(1)	4
Total	4	5	6	3	1	1	20

Problem: The data given below relate to the height and weight of 20 persons. Construct a bivariate frequency table with class interval of height as 62-64, 64-66... and weight as 115-125, 125-135, Write down the marginal distribution of X and Y.

S.No.	Height	Weight	S.No.	Height	Weight
1	70	170	11	70	163
2	65	135	12	67	139
3	65	136	13	63	122
4	64	137	14	68	134
5	69	148	15	67	140
6	63	121	16	69	132
7	65	117	17	65	120
8	70	128	18	68	148
9	71	143	19	67	129
10	62	129	20	67	152

Solution:

Bivariate frequency table showing height and weight of persons.

Height (x) Weight (y)	62-64	64-66	66-68	68-70	70-72	Total
115-125	II (2)	II (2)				4
125-135	I (1)		I (1)	II (2)	I (1)	5
135-145		III (3)	II (2)		I (1)	6
145-155			I (1)	II (2)		3
155-165					I (1)	1
165-175					I (1)	1
Total	3	5	4	4	4	20

The marginal distribution of height and weight are given in the following table.

Marginal distribution of height (X)		Marginal distribution of Weight (Y)	
C1	Frequency	C1	Frequency
62-64	3	115-125	4
64-66	5	125-135	5
66-68	4	135-145	6
68-70	4	145-155	3
70-72	4	155-165	1
Total	20	165-175	1
		Total	20

25 values of two variables X and Y are given below. Form a two-way frequency table showing the relationship between the two. Take class interval of X as 10-20,20-30,..... and Y as 100-200,200-300,.....

X	Y	X	Y	X	Y
12	140	36	315	57	416
24	256	27	440	44	380
33	360	57	390	48	492
22	470	21	590	48	370
44	470	51	250	52	312
37	380	27	550	41	330
29	280	42	360	69	590
55	420	43	570		
48	390	52	290		

Diagrammatic and Graphical Representation

- Classification and Tabulation that help in summarizing the collected data and presenting them in a systematic manner.
- These forms of presentation do not always prove to be interesting to the common man.
- Most convincing and appealing ways in which statistical results may be presented is through diagrams and graphs.
- Just one diagram is enough to represent a given data more effectively than thousand words.

- A layman who has nothing to do with numbers can also understand diagrams.
- Evidence of this can be found in newspapers, magazines, journals, advertisement, etc.

Diagrams

A diagram is a visual form for presentation of statistical data, highlighting their basic facts and relationship.

If we draw diagrams on the basis of the data collected they will easily be understood and appreciated by all.

It is readily intelligible and save a considerable amount of time and energy.

Significance of Diagrams and Graphs

Diagrams and graphs are extremely useful because of the following reasons.

1. They are attractive and impressive.
2. They make data simple and intelligible.
3. They make comparison possible
4. They save time and labour.
5. They have universal utility.
6. They give more information.
7. They have a great memorizing effect.

General rules for constructing diagrams

The diagrammatic presentation of statistical facts will be advantageous provided the following rules are observed in drawing diagrams

1. A diagram should be neatly drawn and attractive.
2. The measurements of geometrical figures used in diagram should be accurate and proportional.
3. The size of the diagrams should match the size of the paper.
4. Every diagram must have a suitable but short heading.
5. The scale should be mentioned in the diagram.

6. Diagrams should be neatly as well as accurately drawn with the help of drawing instruments.
7. Index must be given for identification so that the reader can easily make out the meaning of the diagram.
8. Footnote must be given at the bottom of the diagram.
9. Economy in cost and energy should be exercised in drawing diagram.

Types of diagrams

1. One-dimensional diagrams
2. Two-dimensional diagrams
3. Three-dimensional diagrams
4. Pictograms and Cartograms

One-dimensional diagrams

- In these diagrams, only one-dimensional measurement, ie. height is used and the width is not considered.
- These diagrams are in the form of bar or line charts and can be classified as
 1. Line Diagram
 2. Simple Diagram
 3. Multiple Bar Diagram
 4. Sub-divided Bar Diagram
 5. Percentage Bar Diagram
 6. Deviation Bar diagram

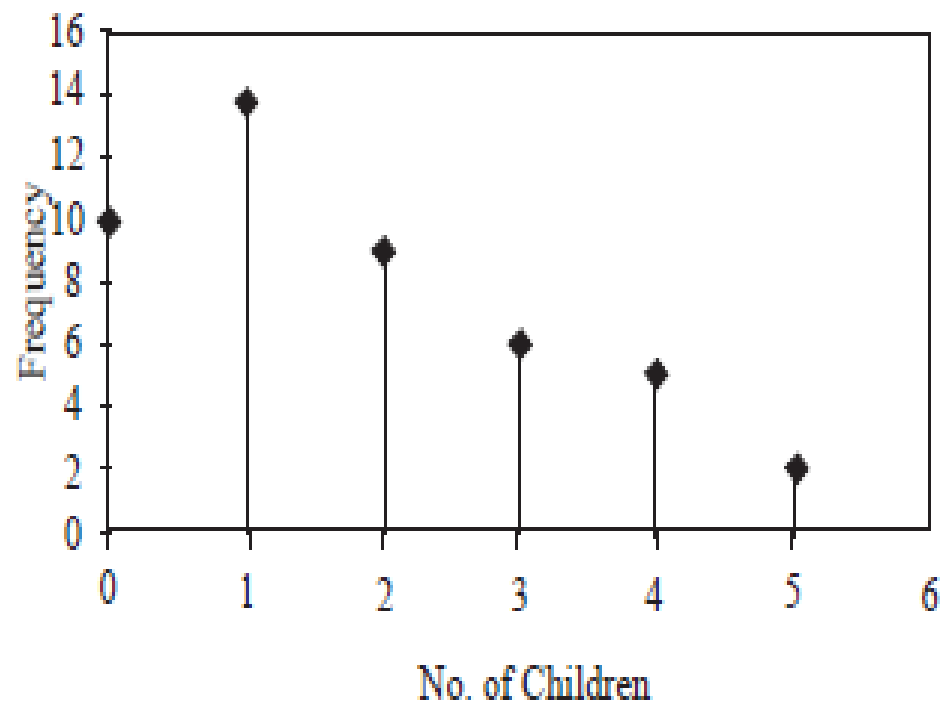
Line Diagram

- Line diagram is used in case where there are many items to be shown and there is not much of difference in their values.
- Such diagram is prepared by drawing a vertical line for each item according to the scale.
- The distance between lines is kept uniform.
- Line diagram makes comparison easy, but it is less attractive

Show the following data by a line chart:

No. of children	0	1	2	3	4	5
Frequency	10	14	9	6	4	2

Line Diagram



Simple Bar Diagram

- Simple bar diagram - drawn either on horizontal or vertical base, but bars on horizontal base more common.
- Bars must be uniform width and intervening space between bars must be equal.
- While constructing a simple bar diagram, the scale is determined on the basis of the highest value in the series.
- To make the diagram attractive, the bars can be coloured. Bar diagram are used in business and economics.

An important limitation of such diagrams is that they can present only one classification or one category of data.

For example, while presenting the population for the last five decades, we can depict the total population in the simple bar diagrams, and not its sex-wise distribution.

Represent the following data by a bar diagram.

Year	Production (in tones)
1991	45
1992	40
1993	42
1994	55
1995	50

Solution :

Simple Bar Diagram

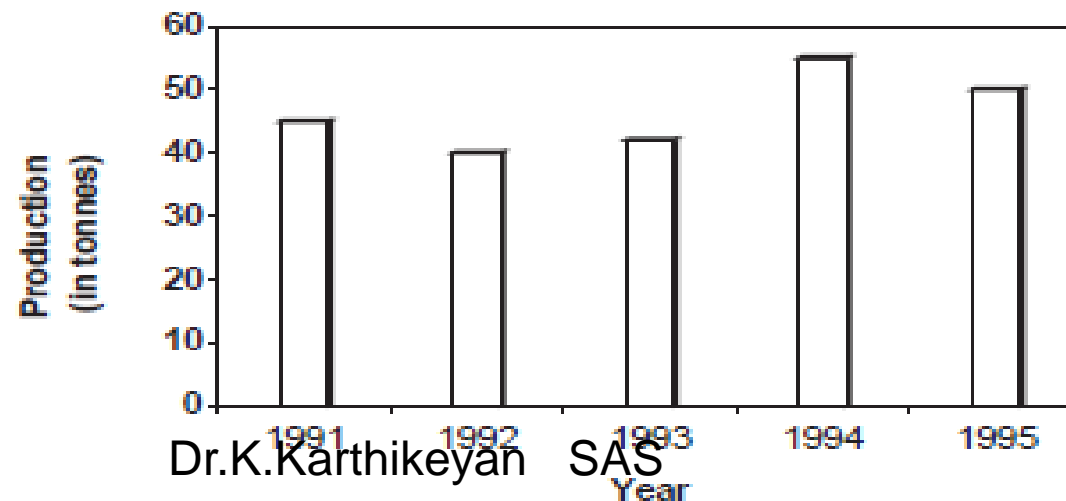


Illustration 1. Represent the following data by a simple bar diagram :

SERVICE TAX COLLECTION (Rs. Crore)

<i>Year</i>	<i>Service Tax</i>
2007-08	51,301
2008-09	60,941
2009-10	58,422
2010-11	71,016
2011-12	97,507
2012-13	1,32,697
2013-14	1,80,141

Solution. The above data is represented by a simple bar diagram.

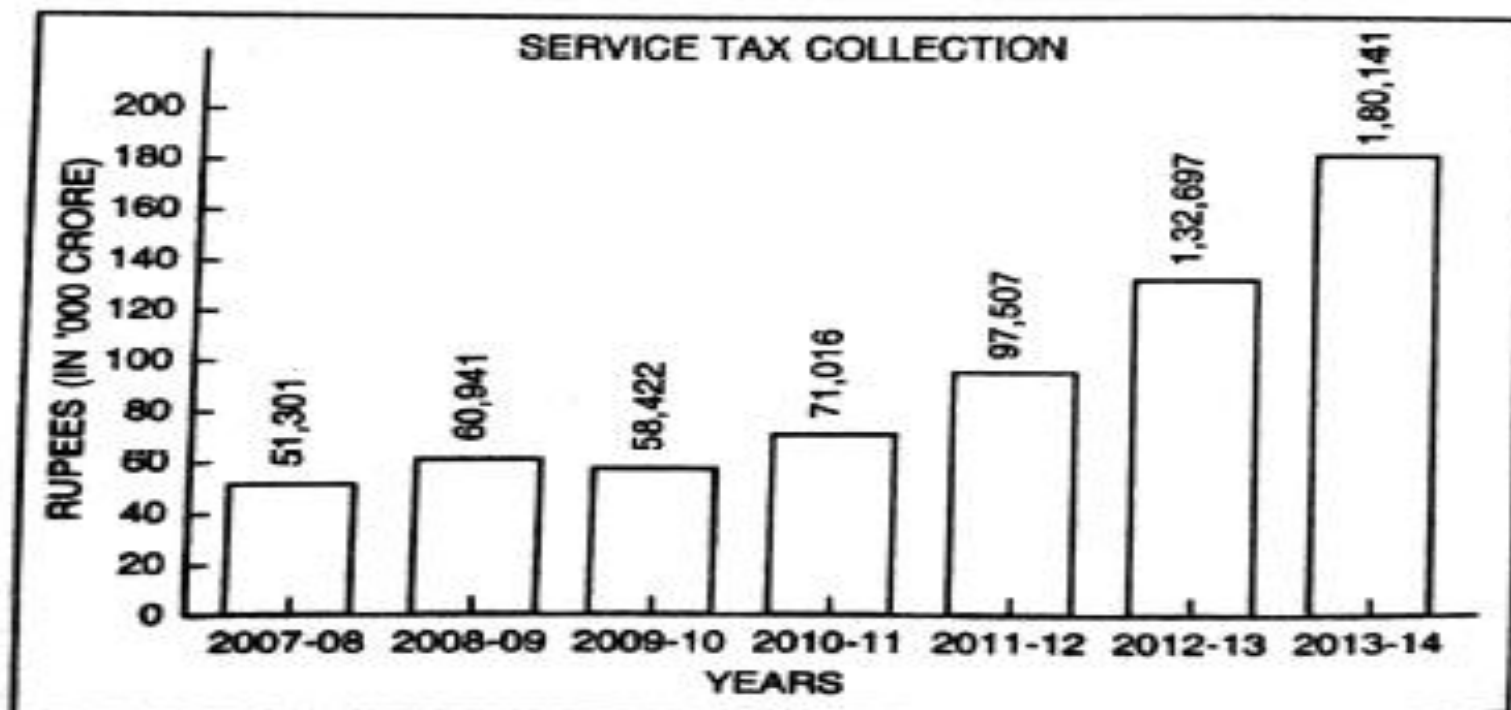
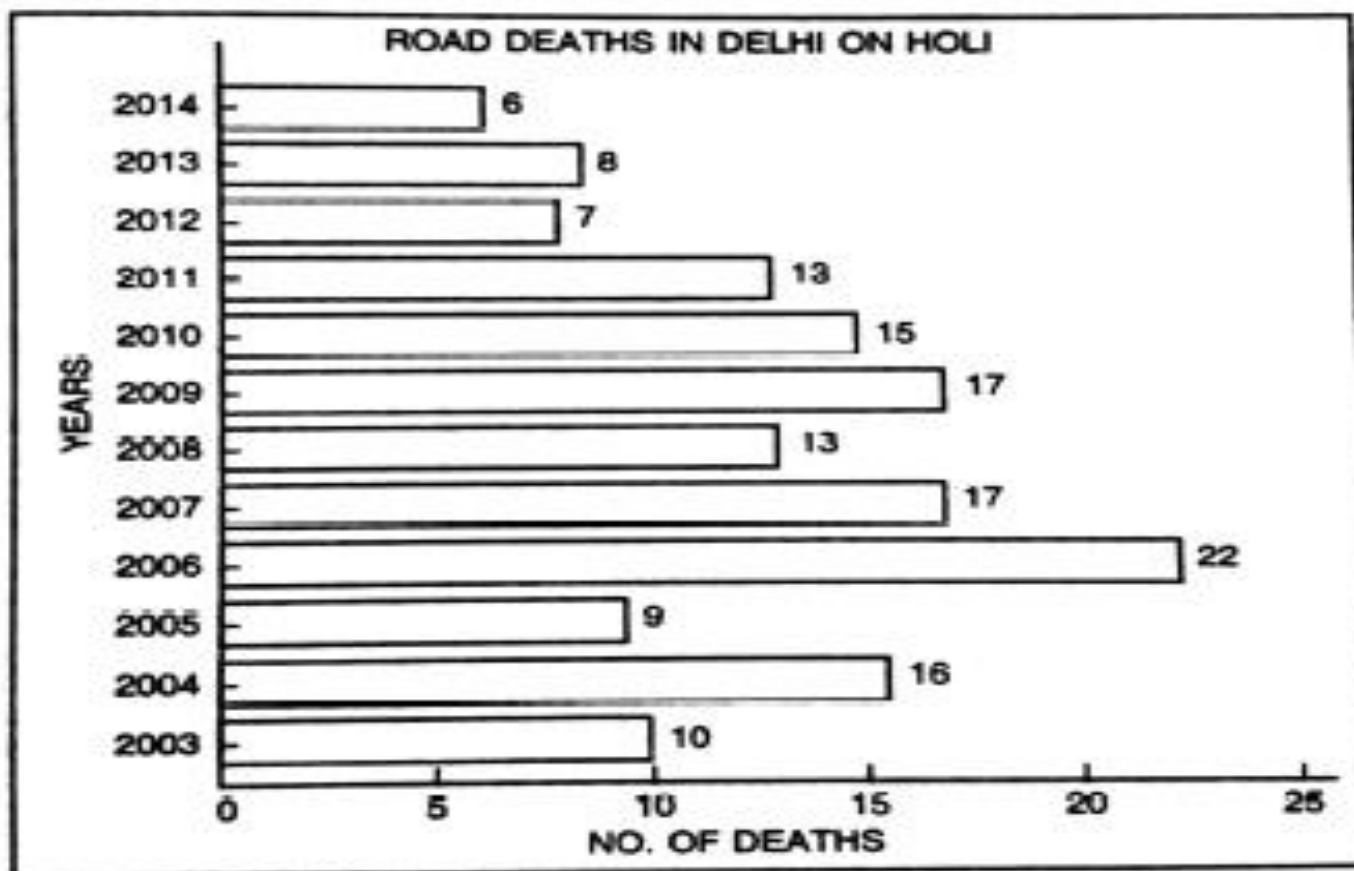


Illustration 3. The following data relates to road deaths in Delhi on the day of Holi from 2003 to 2014 :

Year	:	2003	2004	2005	2006	2007	2008
No. of deaths	:	10	16	9	22	17	13
Year	:	2009	2010	2011	2012	2013	2014
No. of deaths	:	17	15	13	7	8	6

Represent the data by a bar diagram drawn on horizontal basis.

Solution. The above data are represented by a horizontal bar diagram :



Multiple Bar Diagram

- Multiple bar diagram is used for comparing two or more sets of statistical data.
- Bars are constructed side by side to represent the set of values for comparison.
- In order to distinguish bars, they may be either differently coloured or there should be different types of crossings or dotting, etc.
- An index is also prepared to identify the meaning of different colours or dottings.

Draw a multiple bar diagram for the following data.

Year	Profit before tax (in lakhs of rupees)	Profit after tax (in lakhs of rupees)
1998	195	80
1999	200	87
2000	165	45
2001	140	32

Solution :

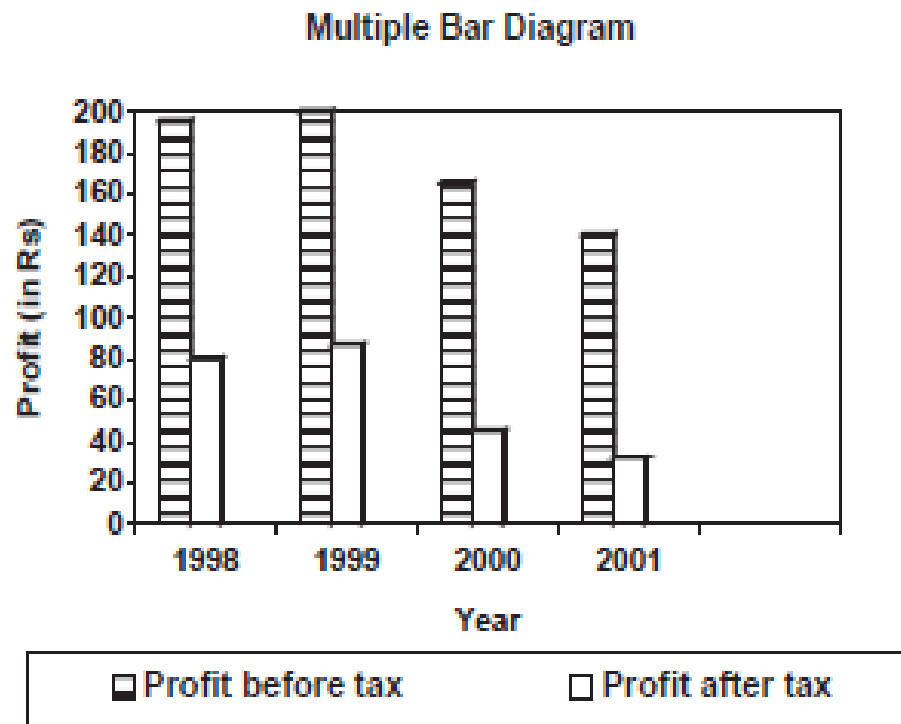
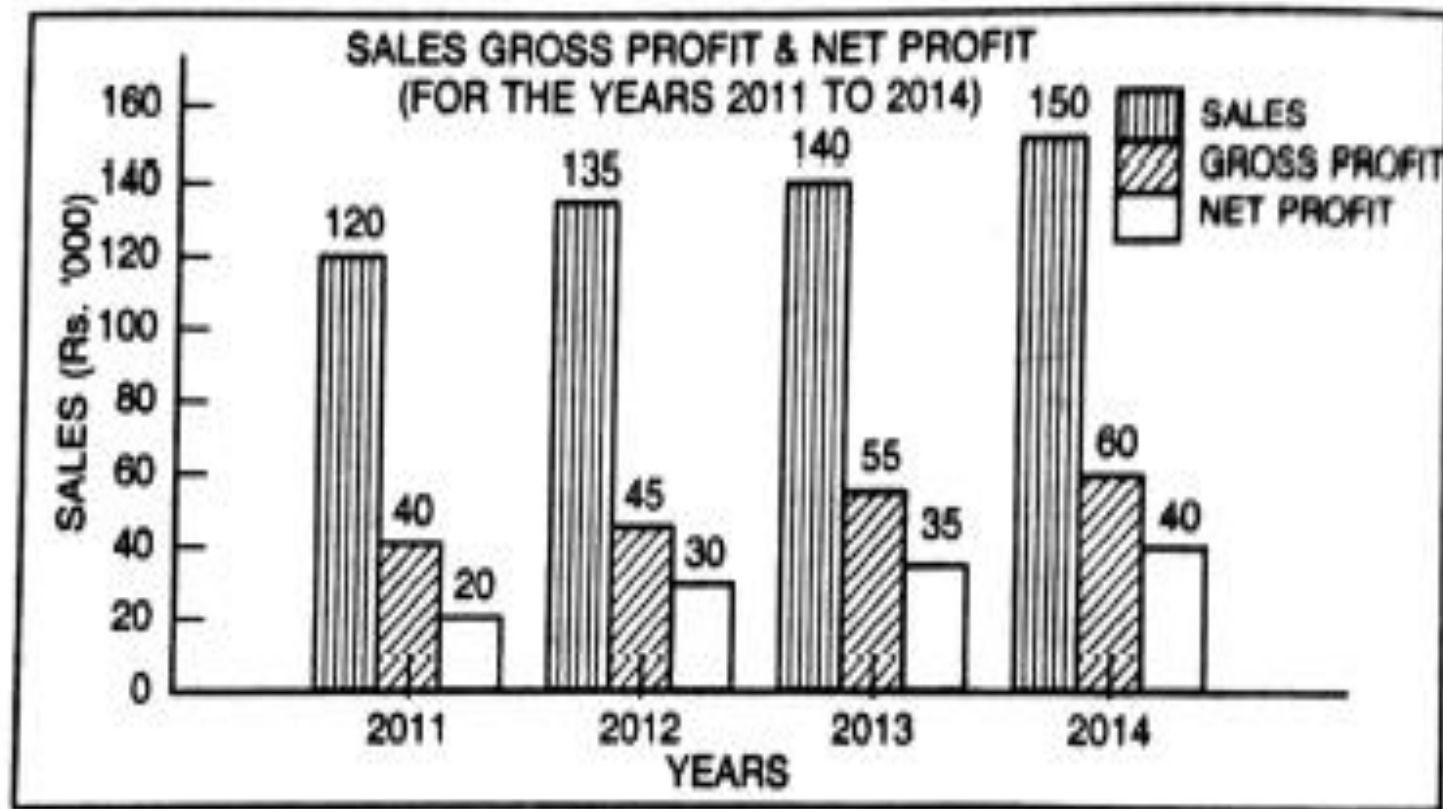


Illustration 7. Draw a multiple bar diagram from the following data:

Year	Sales (Rs.'000)	Gross Profit (Rs.'000)	Net Profit (Rs.'000)
2011	120	40	20
2012	135	45	30
2013	140	55	35
2014	150	60	40

Solution. The above data can be represented by a multiple bar diagram as follows :



Sub-divided Bar Diagram

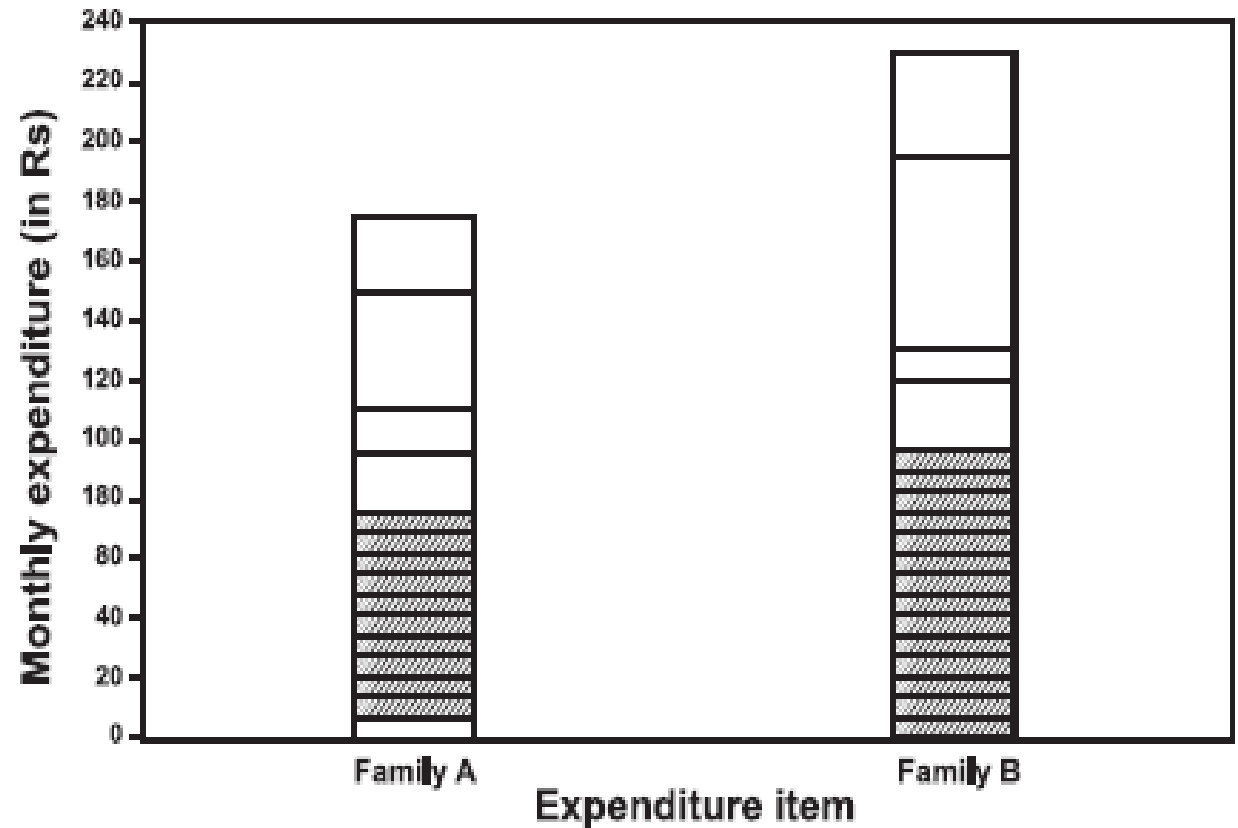
- In a sub-divided bar diagram, the bar is sub-divided into various parts in proportion to the values given in the data and the whole bar represent the total.
- Such diagrams are also called Component Bar diagrams.
- The sub divisions are distinguished by different colours or crossings or dottings.
- The main defect of such a diagram is that all the parts do not have a common base to enable one to compare accurately the various components of the data.

Represent the following data by a sub-divided bar diagram.

Expenditure items	Monthly expenditure (in Rs.)	
	Family A	Family B
Food	75	95
Clothing	20	25
Education	15	10
Housing Rent	40	65
Miscellaneous	25	35

Solution :

Sub-divided Bar Diagram



 Food	 Clothing	 Education
 Housing Rent	 Miscellaneous	

Percentage bar diagram

- Another form of component bar diagram. Here the components are not the actual values but percentages of the whole.
- In the sub-divided bar diagram the bars are of different heights since their totals may be different
- In the percentage bar diagram the bars are of equal height since each bar represents 100 percent.
- In the case of data having sub-division, percentage bar diagram will be more appealing than sub-divided bar diagram

Represent the following data by a percentage bar diagram.

Particular	Factory X	Factory Y
Selling Price	400	650
Quantity Sold	240	365
Wages	3500	5000
Materials	2100	3500
Miscellaneous	1400	2100

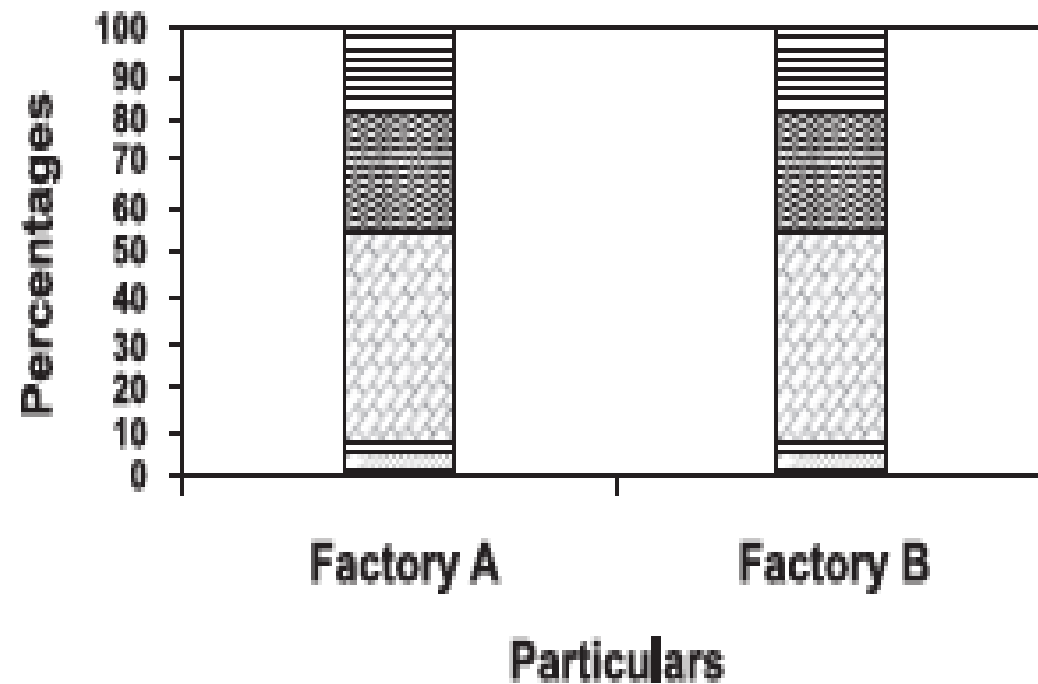
Solution:

Convert the given values into percentages as follows:

Particulars	Factory A		Factory B	
	Rs.	%	Rs.	%
Selling Price	400	5	650	6
Quantity Sold	240	3	365	3
Wages	3500	46	5000	43
Materials	2100	28	3500	30
Miscellaneous	1400	18	2100	18
Total	7640	100	11615	100

Solution :

Sub-divided Percentage Bar Diagram



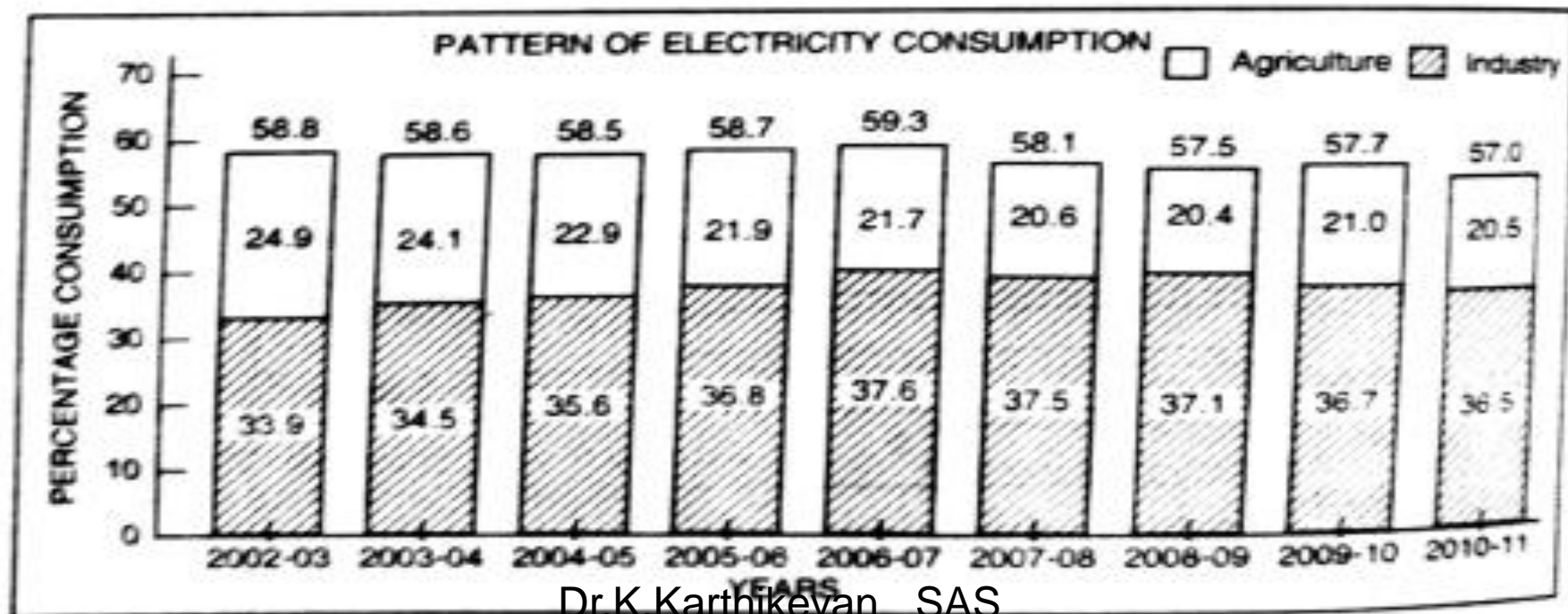
Selling price	Quantity sold
Materials	Miscellaneous

Illustration 4. The following table gives the pattern of Electricity consumption** :

Year	Industry	(Percentage) Agriculture	Total
2002-03	33.9	24.9	58.8
2003-04	34.5	24.1	58.6
2004-05	35.6	22.9	58.5
2005-06	36.8	21.9	58.7
2006-07	37.6	21.7	59.3
2007-08	37.5	20.6	58.1
2008-09	37.1	20.4	57.5
2009-10	36.7	21.0	57.7
2010-11	36.5	20.5	57.0

Represent the data by a suitable diagram.

Solution. The above data can be represented by a sub-divided bar diagram (drawn on a vertical base).

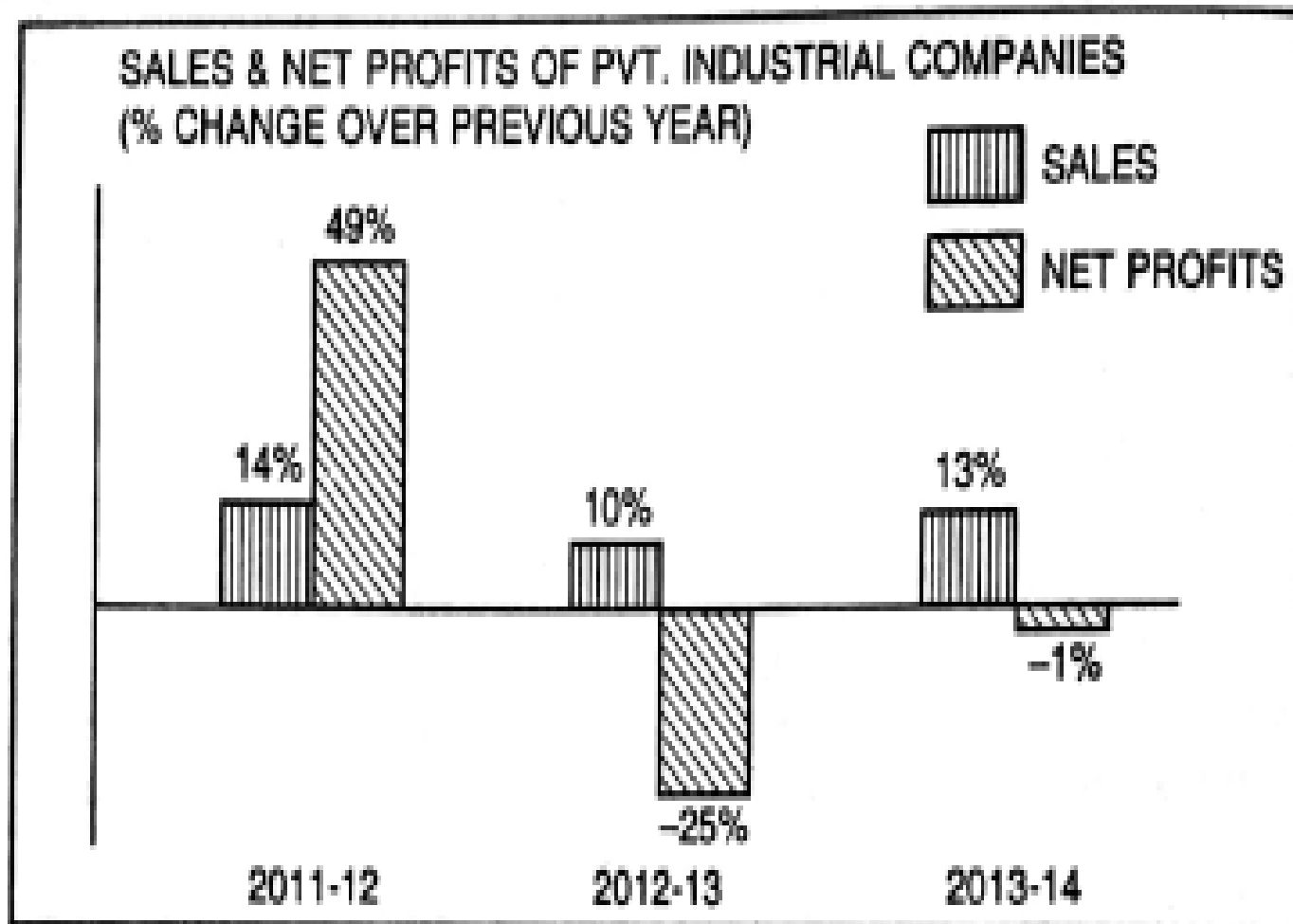


(e) Deviation Bars Deviation bars are popularly used for representing net quantities—excess or deficit, i.e., net profit, net loss, net exports or imports, etc. Such bars can have both positive and negative values. Positive values are shown above the base line and negative values below it. The following illustration would explain this type of diagram:

Illustration 9. Present the following data by a suitable diagram showing the Sales and Net Profits of private industrial companies:

<i>Year</i>	<i>Sales</i>	<i>Net Profits</i>
2011-12	14%	49%
2012-13	10%	-25%
2013-14	13%	-1%

Solution.



RES7001- RM - DATA ANALYSIS

Module – 5

Dr.K.Karthikeyan

**Professor, Dept. of Maths,
SAS, VIT Vellore**

CHI-SQUARE TEST (χ^2)

- F, t and Z tests are based on the assumption that the samples were drawn from normally distributed populations.
- The testing procedure requires assumption about the type of population or parameters, and these tests are known as 'parametric tests'.
- There are many situations in which it is not possible to make any rigid assumption about the distribution of the population from which samples are being drawn. Such type of test called as Non-parametric test.
- Chi-square test of independence and goodness of fit is a prominent example of the use of non-parametric tests
- The square of standard normal variable is known as a chi-square variable with 1 degree of freedom (d.f.).

Properties of Chi-square distribution

- It is a continuous distribution.
- The distribution has only one parameter *i.e.* n d.f.
- The shape of the distribution depends upon the d.f, n .
- The mean of the chi-square distribution is n and variance $2n$
- If U and V are independent random variables having χ^2 distributions with degree of freedom n_1 and n_2 respectively, then their sum $U + V$ has the same χ^2 distribution with d.f $n_1 + n_2$.

The chi-square test is applicable in large number of problems and the test is used to

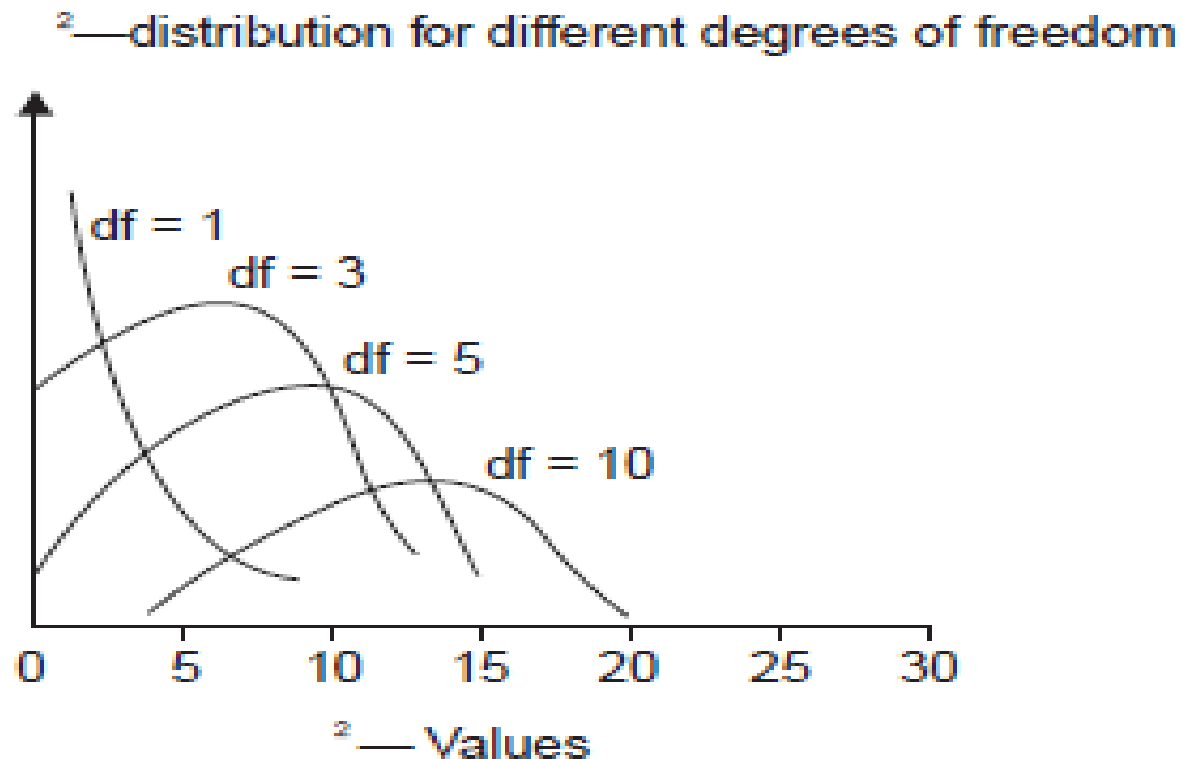
- Test the goodness of fit
- Test the significance of association between two attributes
- Test the homogeneity or the significance of population variance

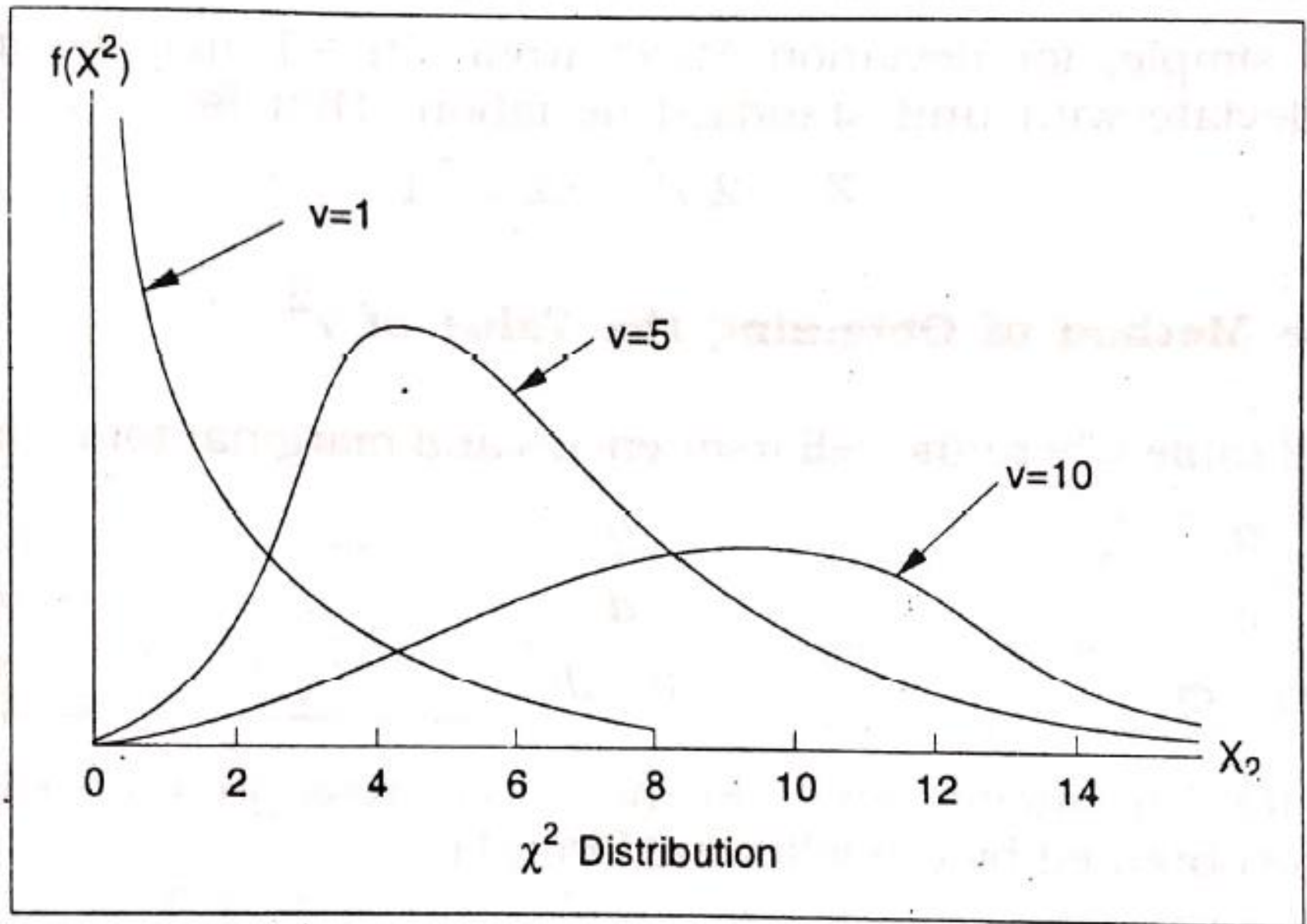
- Chi-square value is used to judge the significance of population variance.
- χ^2 -distribution with n-1 degrees of freedom is

$$\frac{\sigma_s^2}{\sigma_p^2} (n - 1) = \frac{\sigma_s^2}{\sigma_p^2} (d.f.)$$

- χ^2 -distribution is not symmetrical and all the values are positive.
- For different degrees of freedom we have different Chi-square curve.

The smaller the number of degrees of freedom, the more skewed is the distribution.





Testing the population variances- χ^2 –distribution

Null hypothesis : $H_0 : \sigma_s^2 = \sigma_p^2$

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2} (n - 1)$$

where σ_s^2 = variance of the sample;

σ_p^2 = variance of the population;

$(n - 1)$ = degrees of freedom, n being the number of items in the sample.

By comparing the calculated value with the table value of χ^2 for $(n - 1)$ degrees of freedom at a given level of significance, we may either accept or reject the null hypothesis.

If the calculated value of χ^2 is less than the table value, the null hypothesis is accepted, but if the calculated value is equal or greater than the table value, the hypothesis is rejected

Illustration 1

Weight of 10 students is as follows:

S.No.	1	2	3	4	5	6	7	8	9	10
Weight (kg.)	38	40	45	53	47	43	55	48	52	49

Can we say that the variance of the distribution of weight of all students from which the above sample of 10 students was drawn is equal to 20 kgs? Test this at 5 per cent and 1 per cent level of significance.

Solution:

<i>S. No.</i>	X_i (Weight in kgs.)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	38	-9	81
2	40	-7	49
3	45	-2	04
4	53	+6	36
5	47	+0	00
6	43	-4	16
7	55	+8	64
8	48	+1	01
9	52	+5	25
10	49	+2	04
$n = 10$	$\Sigma X_i = 470$	$\Sigma (X_i - \bar{X})^2 = 280$	

$$\bar{X} = \frac{\sum X_i}{n} = \frac{470}{10} = 47 \text{ kgs.}$$

$$\therefore \sigma_s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}} = \sqrt{\frac{280}{10 - 1}} = \sqrt{31.11}$$

$$\text{or } \sigma_s^2 = 31.11$$

Let the null hypothesis be $H_0: \sigma_p^2 = \sigma_s^2$. In order to test this hypothesis we work out the χ^2 value as under:

$$\begin{aligned} \chi^2 &= \frac{\sigma_s^2}{\sigma_p^2} (n - 1) \\ &= \frac{31.11}{20} (10 - 1) = 13.999. \end{aligned}$$

Degrees of freedom in the given case is $(n - 1) = (10 - 1) = 9$.

At 5 per cent level of significance the table value of $\chi^2 = 16.92$ and at 1 per cent level of significance, it is 21.67 for 9 d.f. and both these values are greater than the calculated value of χ^2 which is 13.999.

Hence we accept the null hypothesis and conclude that the variance of the given distribution can be taken as 20 kgs at 5 percent as also at 1 per cent level of significance.

The sample can be said to have been taken from a population with variance 20 kgs.

Illustration 2: A sample of 10 is drawn randomly from a certain population. The sum of the squared deviations from the mean of the given sample is 50. Test the hypothesis that the variance of the population is 5 at 5 per cent level of significance.

Solution: Given information is

$$n = 10$$

$$\Sigma(X_i - \bar{X})^2 = 50$$

$$\therefore s^2 = \frac{\Sigma(X_i - \bar{X})^2}{n - 1} = \frac{50}{9}$$

Take the null hypothesis as $H_0: \sigma_p^2 = \sigma_s^2$. In order to test this hypothesis, we work out the χ^2 value as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n - 1) = \frac{\frac{50}{9}}{5}(10 - 1) = \frac{50}{9} \times \frac{1}{5} \times \frac{9}{1} = 10$$

Degrees of freedom = $(10 - 1) = 9$.

The table value of χ^2 at 5 per cent level for 9 d.f is 16.92. The calculated value of χ^2 is less than this table value, so we accept the null hypothesis and conclude that the variance of the population is 5 as given in the question.

A normal population has mean μ (unknown) and variance 0.018. A random sample of size 20 observations has been taken and its variance is found to be 0.024. Test the null hypothesis $H_0: \sigma^2 = 0.018$ against $H_1: \sigma^2 < 0.018$ at 5% level of significance.

Solution:

Step 1 : **Null Hypothesis** $H_0: \sigma^2 = 0.018$.

i.e. Population variance regarded as 0.018.

Alternative hypothesis $H_1: \sigma^2 < 0.018$.

i.e. Population variance is regarded as less than 0.018.

Step 2 : **Data**

Sample size (n) = 20

Sample variance (s^2) = 0.024

Step 3 : Level of significance

$$\alpha = 5\%$$

Step 4 : Test statistic

Under null hypothesis H_0

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

follows chi-square distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculation of test statistic

The value of chi-square under H_0 is calculated as

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{19 \times 0.024}{0.018} = 25.3$$

Step 6 : Critical value

Since H_1 is a one-sided alternative, the critical values at $\alpha = 0.05$ is $\chi_e^2 = \chi_{19,0.95}^2 = 10.117$.

Step 7 : Decision

Since it is a one-tailed test, the elements of the critical region are determined by the rejection rule $\chi_0^2 < \chi_e^2$

For the given sample information, the rejection rule does not hold, since

$$\chi_0^2 = 25.3 > \chi_e^2 = \chi_{19,0.95}^2 = 10.117.$$

Hence, H_0 is not rejected in favour of H_1 . Thus, the population variance can be regarded as 0.018.

CONDITIONS FOR THE APPLICATION OF χ^2 TEST

The following conditions should be satisfied before χ^2 test can be applied:

- (i) Observations recorded and used are collected on a random basis.
- (ii) All the items in the sample must be independent.
- (iii) No group should contain very few items, say less than 10. In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.
- (iv) The overall number of items must also be reasonably large. It should normally be at least 50, howsoever small the number of groups may be.
- (v) The constraints must be linear. Constraints which involve linear equations in the cell frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known as linear constraints.

CHI-SQUARE AS A NON-PARAMETRIC TEST

Chi-square is an important non-parametric test and as such no rigid assumptions are necessary in respect of the type of population and need the degrees of freedom for using this test.

As a non-parametric test, chi-square can be used for

- (i) Test of goodness of fit
- (ii) Test of independence

The formula to compute Chi square value is

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

O_{ij} = observed frequency of the cell in i th row and j th column.

E_{ij} = expected frequency of the cell in i th row and j th column.

7

Constants of χ^2 Distribution

1. The mean of the χ^2 distribution is equal to the number of degrees of freedom, i.e., $\chi = v$.

2. The variance of the χ^2 distribution is twice the degrees of freedom.
Variance = $2v$.

$$3. \mu_1 = 0, \mu_2 = 2v, \mu_3 = 8v, \mu_4 = 48v + 12v^2.$$

$$4. \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{64v^2}{8v^3} = \frac{8}{v}.$$

$$5. \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{48v + 12v^2}{4v^2} = 3 + \frac{12}{v}.$$

Chi square test of goodness of fit

χ^2 test enables us to see how well does the assumed theoretical distribution (such as Binomial distribution, Poisson distribution or Normal distribution) fit to the observed data.

When some theoretical distribution is fitted to the given data, we are interested in knowing as to how well this distribution fits with the observed data by using the chi-square test.

Computational steps for testing the significance of goodness of fit:

Step 1 : Framing of hypothesis

Null hypothesis H_0 : The goodness of fit is appropriate for the given data set

Alternative hypothesis H_1 : The goodness of fit is not appropriate for the given data set

Step 2 : Data

Calculate the expected frequencies (E_i) using appropriate theoretical distribution such as Binomial or Poisson.

Step 3 : Select the desired level of significance α

Step 4 : Test statistic

The test statistic is

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

where k = number of classes

O_i and E_i are respectively the observed and expected frequency of i^{th} class such that

$$\sum_{i=1}^k O_i = \sum_{i=1}^k E_i .$$

If any of E_i is found less than 5, the corresponding class frequency may be pooled with preceding or succeeding classes such that E_i 's of all classes are greater than or equal to 5. It may be noted that the value of k may be determined after pooling the classes.

The approximate sampling distribution of the test statistic under H_0 is the chi-square distribution with $k-1-s$ d.f, s being the number of parametres to be estimated.

Step 5 : Calculation

Calculate the value of chi-square as

$$\chi_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

The above steps in calculating the chi-square can be summarized in the form of the table as follows:

Step 6 : Critical value

The critical value is obtained from the table of χ^2 for a given level of significance α .

Step 7 : Decision

Decide on rejecting or not rejecting the null hypothesis by comparing the calculated value of the test statistic with the table value, at the desired level of significance.

A die is thrown 132 times with following results:

Number turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Is the die unbiased?

Solution: Let us take the hypothesis that the die is unbiased. If that is so, the probability of obtaining any one of the six numbers is $1/6$ and as such the expected frequency of any one number coming upward is $132 \times 1/6 = 22$. Now we can write the observed frequencies along with expected frequencies and work out the value of χ^2 as follows:

No. turned up	Observed frequency O_i	Expected frequency E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
1	16	22	-6	36	36/22
2	20	22	-2	4	4/22
3	25	22	3	9	9/22
4	14	22	-8	64	64/22
5	29	22	7	49	49/22
6	28	22	6	36	36/22

$$\therefore \sum [(O_i - E_i)^2 / E_i] = 9.$$

Hence, the calculated value of $\chi^2 = 9$.

\therefore Degrees of freedom in the given problem is

$$(n - 1) = (6 - 1) = 5.$$

The table value* of χ^2 for 5 degrees of freedom at 5 per cent level of significance is 11.071.

Comparing calculated and table values of χ^2 , we find that calculated value is less than the table value and as such could have arisen due to fluctuations of sampling. The result, thus, supports the hypothesis and it can be concluded that the die is unbiased.

Five coins are tossed 640 times and the following results were obtained.

Number of heads	0	1	2	3	4	5
Frequency	19	99	197	198	105	22

Fit binomial distribution to the above data.

Solution:

Step 1 : **Null hypothesis** H_0 : Fitting of binomial distribution is appropriate for the given data.

Alternative hypothesis H_1 : Fitting of binomial distribution is not appropriate to the given data.

Step 2 : **Data**

Compute the expected frequencies:

n = number of coins tossed at a time = 5

Let X denote the number of heads (success) in n tosses

N = number of times experiment is repeated = 640

To find mean of the distribution

x	f	fx
0	19	0
1	99	99
2	197	394
3	198	594
4	105	420
5	22	110
Total	640	1617

$$\text{Mean : } \bar{x} = \frac{\sum fx}{\sum f} = \frac{1617}{640} = 2.526$$

The probability mass function of binomial distribution is :

$$p(x) = {}^nC_x p^x q^{n-x}, x = 0, 1, \dots, n$$

Mean of the binomial distribution is $\bar{x} = np$.

Hence,

$$\hat{p} = \frac{\bar{x}}{n} = \frac{2.526}{5} \approx 0.5$$

$$\hat{q} = 1 - \hat{p} \approx 0.5$$

For $x = 0$, the equation (2.1) becomes

$$P(X = 0) = P(0) = {}^5C_0 (0.5)^5 = 0.03125$$

The expected frequency at $x = N P(x)$

The expected frequency at $x=0 : N \times P(0)$

$$= 640 \times 0.03125 = 20$$

We use recurrence formula to find the other expected frequencies.

The expected frequency at $x+1$ is

$$\frac{n-x}{x+1} \left(\frac{p}{q} \right) \times \text{Expected frequency at } x$$

x	$\frac{n-x}{x+1}$	$\frac{p}{q}$	$\frac{n-x}{x+1} \left(\frac{p}{q} \right)$	Expected frequency at $x = N P(x)$
0	5	1	5	20
1	2	1	2	100
2	1	1	1	200
3	0.5	1	0.5	200
4	0.2	1	0.2	100
5	0	1	0	20

Table of expected frequencies:

Number of heads	0	1	2	3	4	5	Total
Expected frequencies	20	100	200	200	100	20	640

Step 3 : Level of significance

$$\alpha = 5\%$$

Step 4 : Test statistic

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Step 5 : Calculation

The test statistic is computed as under:

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
19	20	-1	1	0.050
99	100	-1	1	0.010
197	200	-3	9	0.045
198	200	-2	4	0.020
105	100	5	25	0.250
22	20	2	4	0.200
			Total	0.575

$$\chi_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$
$$= 0.575$$

Step 6 : Critical value

$$\text{Degrees of freedom} = k - 1 - s = 6 - 1 - 1 = 4$$

Critical value for *d.f* 4 at 5% level of significance is 9.488 *i.e.*, $\chi_{4,0.05}^2 = 9.488$

Step 7 : Decision

As the calculated $\chi_0^2 (=0.575)$ is less than the critical value $\chi_{4,0.05}^2 = 9.488$, we do not reject the null hypothesis. Hence, the fitting of binomial distribution is appropriate.

A sample 800 students appeared for a competitive examination. It was found that 320 students have failed, 270 have secured a third grade, 190 have secured a second grade and the remaining students qualified in first grade. The general opinion that the above grades are in the ratio 4:3:2:1 respectively. Test the hypothesis the general opinion about the grades is appropriate at 5% level of significance.

Step 1 : **Null hypothesis** H_0 : The result in four grades follows the ratio 4:3:2:1

Alternative hypothesis H_1 : The result in four grades does not follows the ratio 4:3:2:1

Step 2 : **Data**

Compute expected frequencies:

Under the assumption on H_0 , the expected frequencies of the four grades are:

$$\frac{4}{10} \times 800 = 320 \quad \frac{3}{10} \times 800 = 240 \quad \frac{2}{10} \times 800 = 160 \quad \frac{1}{10} \times 800 = 80$$

Step 3 : Test statistic

The test statistic is computed using the following table.

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
320	320	0	0	0
270	240	30	900	3.75
190	160	30	900	5.625
20	80	-60	3600	45
			Total	54.375

The test statistic is calculated as

$$\begin{aligned}\chi_0^2 &= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \\ &= 54.375\end{aligned}$$

Step 4 : Critical value

The critical value of χ^2 for 3 d.f. at 5% level of significance is 7.81 i.e., $\chi_{3,0.05}^2 = 7.81$

Step 5 : Decision

As the calculated value of $\chi_0^2 (=54.375)$ is greater than the critical value $\chi_{3,0.05}^2 = 7.81$, reject H_0 . Hence, the results of the four grades do not follow the ratio 4:3:2:1.

A company keeps records of accidents. During a recent safety review, a random sample of 60 accidents was selected and classified by the day of the week on which they occurred.

Day	:	MON	TUE	WED	THU	FRI
No. of accidents	:	8	12	9	14	17

Test whether there is any evidence that accidents are more likely on some days than others.

Solutions:

H_0 : Accidents are equally likely to occur on any day of the week.

H_1 : Accidents are not equally likely to occur on the days of the week.

Total number of accidents = 60

On the assumption H_0 , the expected number of accidents on any day

$$= \frac{60}{5} = 12$$

Let O denote observed frequency and E denote expected frequency

O	E	O-E	$(O-E)^2$
8	12	-4	16
12	12	0	0
9	12	-3	9
14	12	2	4
17	12	5	25
60	60		54

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right] = \frac{54}{12} = 4.5$$

n = number of classes = 5

\therefore number of degrees of freedom = $n - 1 = 5 - 1 = 4$

For 4 of degrees of freedom the table value of $\chi^2 = 9.488$.

But the calculated value of χ^2 is 4.5.

\therefore Calculated value of $\chi^2 <$ the table value of χ^2 .

Hence H_0 is accepted at 5% level. This means that the accidents are equally likely to occur on any day of the week.

An insurance company, which specialises in motor insurance, bases the premium charged partly on the region of the country in which the owner of the vehicle lives. The following table shows the number of claims from each region for a random sample of claims in the year 1990 and the percentage distribution of claims in the previous years.

Region	:	A	B	C	D	E
Number of claims in 1990	:	59	119	67	161	94
% of claims before 1990	:	10	25	15	30	20

Test whether there is evidence of a significant change in the distribution of claims.

Solution:

H_0 : There is no evidence of significant change in the distribution of claims,

H_1 : There is evidence of significant change in the distribution of claims,

Total observed frequency = 500

\therefore On the assumption H_0 , the expected frequencies for the different classes

are $\frac{10}{100} \times 500, \frac{25}{100} \times 500, \frac{15}{100} \times 500, \frac{30}{100} \times 500, \frac{20}{100} \times 500$

i.e., 50, 125, 75, 150, 100.

Region	O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
A	59	50	9	81	1.620
B	119	125	-6	36	0.228
C	67	75	-8	64	0.853
D	161	150	11	121	0.807
E	94	100	-6	36	0.360
	500				3.928

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E} \right] = 3.928$$

Number of degrees of freedom = $5 - 1 = 4$

Table value for 4 df at 5% level is 9.48,

Conclusion:

Since the calculated value of χ^2 is less than the table value, H_0 is accepted at 5% level. (i.e.) there is no evidence of significant change in the distribution of claims.

The theory predicts that the proportion of beans in four given groups should be $9 : 3 : 3 : 1$. In an examination with 1600 beans, the numbers in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory?

Solution:

H_0 : The proportion of beans in the four groups are in the ratio $9 : 3 : 3 : 1$.

H_1 : The proportion of beans in the four groups are not in the ratio $9 : 3 : 3 : 1$.

On the assumption H_0 , the expected frequencies are

$$\frac{9}{16} \times 1600, \frac{3}{16} \times 1600, \frac{3}{16} \times 1600, \text{ and } \frac{1}{16} \times 1600$$

(i.e.) 900, 300, 300, 100.

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
882	900	-18	324	0.360
313	300	13	169	0.563
287	300	-13	169	0.563
118	100	18	324	3.240
	1600			4.726

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right] = 4.726$$

Number of degrees of freedom $= 4 - 1 = 3$

Table value of $\chi^2 = 7.81$.

Conclusion:

The calculated value of χ^2 is less than the table value of χ^2 .

$\therefore H_0$ is accepted at 5% level.

\therefore The beans in the four groups are in the ratio 9 : 3 : 3 : 1.

χ^2 -Test of Independence of Attributes

If the population is known to have two major attributes A and B , then A can be divided into m categories A_1, A_2, \dots, A_m and B can be divided into n categories B_1, B_2, \dots, B_n . Accordingly the members of the population and hence those of the sample can be divided into mn classes. In this case, the sample data may be presented in the form of a matrix containing m rows and n columns and hence mn cells and showing the observed frequencies O_{ij} , in the various cells, where $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$. O_{ij} means the number of observed frequencies possessing the attributes A_i and B_j . The matrix or tabular form of the sample data, called an $(m \times n)$ contingency table is given below:

$A \setminus B$	B_1	B_2	-	B_j	-	B_n	Row Total
A_1	O_{11}	O_{12}	-	O_{1j}	-	O_{1n}	O_{1*}
A_2	O_{21}	O_{22}	-	O_{2j}	-	O_{2n}	O_{2*}
\vdots	-	-	-	-	-	-	-
A_i	O_{i1}	O_{i2}	-	O_{ij}	-	O_{in}	O_{i*}
\vdots	-	-	-	-	-	-	-
A_m	O_{m1}	O_{m2}	-	O_{mj}	-	O_{mn}	O_{m*}
Column Total	O_{*1}	O_{*2}	-	O_{*j}	-	O_{*n}	N

Now, based on the null hypothesis H_0 i.e. the assumption that the two attributes A and B are independent, we compute the expected frequencies E_{ij} for

various cells, using the following formula $E_{ij} = \frac{O_{i*} \cdot O_{*j}}{N}$, $i = 1, 2, \dots, m$; and

$j = 1, 2, \dots, n$

i.e. $E_{ij} = \left\{ \frac{\left(\text{Total of observed frequencies in the } i^{\text{th}} \text{ row} \right) \times \left(\text{total of observed frequencies in the } j^{\text{th}} \text{ column} \right)}{\text{Total of all cell frequencies}} \right\}$

Then we compute $\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \left\{ \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right\}$

The number of degrees of freedom for this χ^2 computed from the $(m \times n)$ contingency table is $v = (m - 1)(n - 1)$.

If $\chi^2 < \chi_v^2(\alpha)$, H_0 is accepted at $\alpha\%$ LOS i.e. the attributes A and B are independent.

If $\chi^2 > \chi_v^2(\alpha)$, H_0 is rejected at $\alpha\%$ LOS i.e. A and B are not independent.

Chi-square test of Independence of attribute

χ^2 test enables us to explain whether or not two attributes are associated

Steps involved in applying chi-square test

- (i) First of all calculate the expected frequencies on the basis of given hypothesis or on the basis of null hypothesis. Usually in case of a 2×2 or any contingency table, the expected frequency for any given cell is worked out as under:

$$\text{Expected frequency of any cell} = \frac{\left[\begin{array}{l} (\text{Row total for the row of that cell}) \times \\ (\text{Column total for the column of that cell}) \end{array} \right]}{(\text{Grand total})}$$

- (ii) Obtain the difference between observed and expected frequencies and find out the squares of such differences i.e., calculate $(O_{ij} - E_{ij})^2$.
- (iii) Divide the quantity $(O_{ij} - E_{ij})^2$ obtained as stated above by the corresponding expected frequency to get $(O_{ij} - E_{ij})^2/E_{ij}$ and this should be done for all the cell frequencies or the group frequencies.

- (iv) Find the summation of $(O_{ij} - E_{ij})^2/E_{ij}$ values or what we call $\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$. This is the required χ^2 value.

The χ^2 value obtained as such should be compared with relevant table value of χ^2 and then inference be drawn as stated above.

χ^2 -Table

<i>n</i>	<i>Probability</i>					
	0.99	0.95	0.10	0.05	0.02	0.01
1	0.000157	0.00393	2.706	3.841	5.412	6.635
2	0.0201	0.103	4.605	5.991	7.824	9.210
3	0.115	0.352	6.251	7.815	9.837	11.345
4	0.297	0.711	7.779	9.488	11.668	13.277
5	0.554	1.145	9.236	11.070	13.388	15.086
6	0.872	1.635	10.645	12.592	15.033	16.812
7	1.238	2.167	12.017	14.067	16.622	18.475
8	1.646	2.733	13.362	15.507	18.168	20.090
9	2.088	3.325	14.684	16.919	19.670	21.666
10	2.558	3.940	15.987	18.307	21.161	23.209
11	3.053	4.575	17.275	19.675	22.618	24.725
12	3.571	5.226	18.549	21.026	24.054	26.217
13	4.107	5.982	19.812	22.362	25.472	27.688
14	4.660	6.571	21.064	23.685	26.873	29.141
15	5.229	7.261	22.307	24.996	28.259	30.578
16	5.812	7.962	23.542	26.296	29.633	32.000
17	6.408	8.672	24.768	27.587	30.995	33.409
18	7.015	9.390	25.989	28.869	32.346	34.805
19	7.633	10.117	27.204	30.114	33.687	36.191
20	8.260	10.851	28.412	31.410	35.020	37.566
21	8.897	11.581	29.615	32.671	36.343	38.932
22	9.542	12.338	30.813	33.924	37.659	40.289
23	10.196	13.091	32.007	35.172	38.968	41.638
24	10.856	13.848	33.196	36.415	40.270	42.980
25	11.524	14.611	34.382	37.652	41.566	44.314
26	12.198	15.379	35.563	38.885	42.856	45.642
27	12.879	16.151	36.741	40.113	44.140	46.963
28	13.565	16.924	37.916	41.337	45.419	48.278
29	14.256	17.708	39.087	42.557	46.693	49.588

The table given below shows the data obtained during outbreak of smallpox:

	<i>Attacked</i>	<i>Not attacked</i>	<i>Total</i>
Vaccinated	31	469	500
Not vaccinated	185	1315	1500
Total	216	1784	2000

Test the effectiveness of vaccination in preventing the attack from smallpox. Test your result with the help of χ^2 at 5 per cent level of significance.

Solution: Let us take the hypothesis that vaccination is not effective in preventing the attack from smallpox i.e., vaccination and attack are independent. On the basis of this hypothesis, the expected frequency corresponding to the number of persons vaccinated and attacked would be:

$$\text{Expectation of } (AB) = \frac{(A) \times (B)}{N}$$

when A represents vaccination and B represents attack.

$$\therefore \begin{aligned} (A) &= 500 \\ (B) &= 216 \\ N &= 2000 \end{aligned}$$

$$\text{Expectation of } (AB) = \frac{500 \times 216}{2000} = 54$$

Now using the expectation of (AB) , we can write the table of expected values as follows:

	<i>Attacked: B</i>	<i>Not attacked: b</i>	<i>Total</i>
Vaccinated: <i>A</i>	$(AB) = 54$	$(Ab) = 446$	500
Not vaccinated: <i>a</i>	$(aB) = 162$	$(ab) = 1338$	1500
Total	216	1784	2000

Group	Observed frequency O_{ij}	Expected frequency E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2$	$(O_{ij} - E_{ij})^2/E_{ij}$
AB	31	54	-23	529	529/54=9.796
Ab	469	446	+23	529	529/44=1.186
aB	158	162	+23	529	529/162=3.265
ab	1315	1338	-23	529	529/1338=0.395

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 14.642$$

\therefore Degrees of freedom in this case $= (r - 1) (c - 1) = (2 - 1) (2 - 1) = 1$.

The table value of χ^2 for 1 degree of freedom at 5 per cent level of significance is 3.841. The calculated value of χ^2 is much higher than this table value and hence the result of the experiment does not support the hypothesis. We can, thus, conclude that vaccination is effective in preventing the attack from smallpox.

In a 2x2 table where the cell frequencies and marginal totals are as below:

a	b	(a+b)
c	d	(c+d)

(a+c)	(b+d)	N
-------	-------	---

N is the total frequency and ad the larger cross-product, the value of χ^2 can easily be obtained by the following formula:

$$\chi^2 = \frac{N (ad - bc)^2}{(a + c) (b + d) (c + d) (a + b)} \quad \text{or}$$

With Yate's corrections

$$\chi^2 = \frac{N (ab - bc - \frac{1}{2}N)^2}{(a + c) (b + d) (c + d) (a + b)}$$

Illustration:

In an anti-diabetes campaign in a certain area, a particular medicine, say x was administered to 812 persons out of a total population of 3248. The number of diabetes cases is shown below:

Treatment	Diabetes	No Diabetes	Total
Medicine x	20	792	812
No Medicine x	220	2216	2436
Total	240	3008	3248

Discuss the usefulness of medicine x in checking malaria.

Solution:

Let us take the hypothesis that quinine is not effective in checking diabetes. Applying χ^2 test :

$$\text{Expectation of (AB)} = \frac{(A) \times (B)}{N} = \frac{240 \times 812}{3248} = 60$$

Or E_1 , i.e., expected frequency corresponding to first row and first column is 60. The table of expected frequencies shall be:

60	752	812
180	2256	2436
240	3008	3248

O	E	$(O - E)^2$	$(O - E)^2/E$
20	60	1600	26.667
220	180	1600	8.889
792	752	1600	2.218
2216	2256	1600	0.709
$[\Sigma(O - E)^2/E] = 38.593$			

$$\chi^2 = [\Sigma(O - E)^2/E] = 38.593$$

$$V = (r - 1) (c - 1) = (2 - 1) (2 - 1) = 1$$

For

$$v = 1, \chi^2_{0.05} = 3.84$$

The calculated value of χ^2 is greater than the table value. The hypothesis is rejected. Hence medicine x is useful in checking malaria.

Example 10. In an industry, 200 workers, employed for a specific job, were classified according to their performance and training received / not received to test independence of a specific training and performance. The data is summarised as follows :

	Performance		Total
	Good	Not Good	
Trained	100	50	150
Untrained	20	30	50
Total	120	80	200

Use χ^2 test of independence at 5% level of significance and write your conclusion.

[Data from χ^2 -table; χ^2 (1 d.f. 5%) = 3.84]

Solution. We have the following table for expected values.

Table: Expected Frequencies

	<i>Performance</i>		<i>Total</i>
	<i>Good</i>	<i>Not Good</i>	
Trained	$\frac{150 \times 120}{200} = 90$	$\frac{150 \times 80}{200} = 60$	150
Untrained	$\frac{120 \times 50}{200} = 30$	$\frac{80 \times 50}{200} = 20$	50
Total	120	80	200

Null Hypotheses H_0 : No association exists between performance and training of workers.

Alternate Hypothesis H_1 : An association exists between performance and training of workers.

$$\begin{aligned}\text{Now } \psi^2 &= \sum \frac{(O - E)^2}{E} = \frac{(100 - 90)^2}{90} + \frac{(50 - 60)^2}{60} + \frac{(20 - 30)^2}{30} + \frac{(30 - 20)^2}{20} \\ &= \frac{100}{90} + \frac{100}{60} + \frac{100}{30} + \frac{100}{20} = \frac{2000}{180} = 11.11.\end{aligned}$$

Degree of freedom: $\nu = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$.

Decision: Since computed value of $\psi^2 (= 11.11)$ at 1 d.f. is greater than critical value of ψ^2 at 5% significance level and 1 d.f., i.e., 3.84 so we reject null hypothesis and accept the alternative hypothesis. Hence an association exists between performance and training workers.

Illustration 7

Two research workers classified some people in income groups on the basis of sampling studies. Their results are as follows:

<i>Investigators</i>	<i>Income groups</i>			<i>Total</i>
	<i>Poor</i>	<i>Middle</i>	<i>Rich</i>	
<i>A</i>	160	30	10	200
<i>B</i>	140	120	40	300
Total	300	150	50	500

Show that the sampling technique of at least one research worker is defective.

Solution: Let us take the hypothesis that the sampling techniques adopted by research workers are similar (i.e., there is no difference between the techniques adopted by research workers). This being so, the expectation of *A* investigator classifying the people in

<i>Groups</i>	<i>Observed frequency</i> O_y	<i>Expected frequency</i> E_y	$O_y - E_y$	$(O_y - E_y)^2 E_y$
<i>Investigator A</i>				
classifies people as poor	160	120	40	1600/120 = 13.33
classifies people as middle class people	30	60	-30	900/60 = 15.00
classifies people as rich	10	20	-10	100/20 = 5.00
<i>Investigator B</i>				
classifies people as poor	140	180	-40	1600/180 = 8.88
classifies people as middle class people	120	90	30	900/90 = 10.00
classifies people as rich	40	30	10	100/30 = 3.33

Hence,

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 55.54$$

$$\begin{aligned} \therefore \text{Degrees of freedom} &= (c - 1)(r - 1) \\ &= (3 - 1)(2 - 1) = 2. \end{aligned}$$

The table value of χ^2 for two degrees of freedom at 5 per cent level of significance is 5.991.

The calculated value of χ^2 is much higher than this table value which means that the calculated value cannot be said to have arisen just because of chance. It is significant. Hence, the hypothesis does not hold good. This means that the sampling techniques adopted by two investigators differ and are not similar. Naturally, then the technique of one must be superior than that of the other.

YATES' CORRECTION

F. Yates has suggested a correction for continuity in χ^2 value calculated in connection with a (2×2) table, particularly when cell frequencies are small (since no cell frequency should be less than 5 in any case, though 10 is better as stated earlier) and χ^2 is just on the significance level. The correction suggested by Yates is popularly known as Yates' correction. It involves the reduction of the deviation of observed from expected frequencies which of course reduces the value of χ^2 . The rule for correction is to adjust the observed frequency in each cell of a (2×2) table in such a way as to reduce the deviation of the observed from the expected frequency for that cell by 0.5, but this adjustment is made in all the cells without disturbing the marginal totals. The formula for finding the value of χ^2 after applying Yates' correction can be stated thus:

$$\chi^2(\text{corrected}) = \frac{N \cdot (|ad - bc| - 0.5N)^2}{(a+b)(c+d)(a+c)(b+d)}$$

In case we use the usual formula for calculating the value of chi-square viz.,

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

then Yates' correction can be applied as under:

$$\chi^2(\text{corrected}) = \frac{[|O_1 - E_1| - 0.5]^2}{E_1} + \frac{[|O_2 - E_2| - 0.5]^2}{E_2} + \dots$$

It may again be emphasised that Yates' correction is made only in case of (2×2) table and that too when cell frequencies are small.

The following information is obtained concerning an investigation of 50 ordinary shops of small size:

	<i>Shops</i>		<i>Total</i>
	<i>In towns</i>	<i>In villages</i>	
Run by men	17	18	35
Run by women	3	12	15
Total	20	30	50

Can it be inferred that shops run by women are relatively more in villages than in towns? Use χ^2 test.

Solution: Take the hypothesis that there is no difference so far as shops run by men and women in towns and villages. With this hypothesis the expectation of shops run by men in towns would be:

$$\text{Expectation of } (AB) = \frac{(A) \times (B)}{N}$$

<i>Groups</i>	<i>Observed frequency</i> O_{ij}	<i>Expected frequency</i> E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2/E_{ij}$
(AB)	17	14	3	9/14=0.64
(Ab)	18	21	-3	9/21=0.43
(aB)	3	6	-3	9/6=1.50
(ab)	12	9	3	9/9=1.00

$$\therefore \chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 3.57$$

As one cell frequency is only 3 in the given 2×2 table, we also work out χ^2 value applying Yates' correction and this is as under:

$$\begin{aligned}\chi^2(\text{corrected}) &= \frac{[|17 - 14| - 0.5]^2}{14} + \frac{[|18 - 21| - 0.5]^2}{21} + \frac{[|3 - 6| - 0.5]^2}{6} + \frac{[|12 - 9| - 0.5]^2}{9} \\&= \frac{(2.5)^2}{14} + \frac{(2.5)^2}{21} + \frac{(2.5)^2}{6} + \frac{(2.5)^2}{9} \\&= 0.446 + 0.298 + 1.040 + 0.694 \\&= 2.478\end{aligned}$$

$$\therefore \text{Degrees of freedom} = (c - 1)(r - 1) = (2 - 1)(2 - 1) = 1$$

\therefore Degrees of freedom $= (c - 1) (r - 1) = (2 - 1) (2 - 1) = 1$

Table value of χ^2 for one degree of freedom at 5 per cent level of significance is 3.841. The calculated value of χ^2 by both methods (i.e., before correction and after Yates' correction) is less than its table value. Hence the hypothesis stands. We can conclude that there is no difference between shops run by men and women in villages and towns.

100 students randomly selected from the 1000 students enrolled on MBA Programme were cross classified by age and grade point. Accordingly, the following data were compiled.

Grade Pt.	Age (in yrs.)			Total
	25 & under	26 – 28	over 28	
upto 3.0	6	9	5	20
3.1 – 3.5	18	14	8	40
3.6 – 4.0	11	12	17	40
Total	35	35	30	100

At 5% level of significance, test the hypothesis that age and grade points are independent.

Solution:

H_0 : Age and grade points are independent.

H_1 : Age and grade points are not independent.

Table showing expected frequencies

Grade Pt.	Age (in yrs.)			Total
	25 & under	26 – 28	over 28	
upto 3.0	$\frac{35 \times 20}{100} = 7$	7	6	20
3.1 – 3.5	14	14	12	40
3.6 – 4.0	14	14	12	40
Total	35	35	30	100

<i>O</i>	<i>E</i>	<i>O-E</i>	$(O-E)^2$	$\frac{(O-E)^2}{E}$
6	7	-1	1	0.143
9	7	2	4	0.571
5	6	-1	1	0.167
18	14	4	16	1.143
14	14	0	0	0.000
8	12	-4	16	1.333
11	14	-3	9	0.643
12	14	-2	4	0.286
17	12	5	25	2.083
100	100			6.369

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E} \right] = 6.369$$

$$\text{ndf} = (3 - 1) \times (3 - 1) = 4$$

Table values of χ^2 for 4 df at 5% level = 9.489

Conclusion:

H_0 is accepted since the calculated value $\chi^2 <$ the table value of χ^2 .
Hence Age and grade points are independent.

Calculate the expected frequencies for the following data presuming that the two attributes viz, condition of home and condition of the child are independent.

		Condition of the Home	
		Clean	Dirty
Condition of the child	Clean	70	50
	Fairly Clean	80	20
	Dirty	35	45

Use χ^2 – test of 5% level to state whether the two attributes are independent.

Solution:

H_0 : Condition of the child and condition of the home are independent.

H_1 : Condition of the child and condition of the home are not independent.

Observed Frequency Table

	Condition of the Home		Total
	Clean	Dirty	
Condition of the child			
Clean	70	50	120
Fairly clean	80	20	100
Dirty	35	45	80
Total	185	115	300

Expected Frequency Table

	Condition of the Home		
	Clean	Dirty	
	$\frac{185 \times 120}{300} = 74$		
Condition of Clean the child		46	120
Fairly clean	62	38	100
Dirty	49	31	80
Total	185	115	300

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
70	74	-4	16	0.22
50	46	4	16	0.35
80	62	18	324	5.23
20	38	-18	324	8.53
35	49	-14	196	4.00
45	31	14	196	6.32
300	300			24.65

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E} \right] = 24.65$$

$$\text{ndf} = (3 - 1) \times (2 - 1) = 2$$

Table value of χ^2 for 2 df at 5% level = 5.991

Conclusion:

H_0 is rejected since the calculated value of $\chi^2 >$ the table value of χ^2 .

Hence the condition of the child and the condition of the home are not independent.

A sample of hotels in a particular country was selected. The following table shows the number of hotels in each region of the country and in each of four grades.

Grade	Region		
	Eastern	Central	Western
1 Star	29	22	29
2 Star	67	38	55
3 Star	53	32	35
4 Star	11	8	21

Show that there is evidence of a significant association between region and grade of hotel in this country.

Table showing observed Frequencies.

Grade	Region			Total
	Eastern	Central	Western	
1 Star	29	22	29	80
2 Star	67	38	55	160
3 Star	53	32	35	120
4 Star	11	8	21	40
Total	160	100	140	400

Table Showing Expected Frequencies.

Grade	Region			Total
	Eastern	Central	Western	
1 Star	32	20	28	80
2 Star	64	60	56	160
3 Star	48	30	42	120
4 Star	16	10	14	40
Total	160	100	140	400

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
29	32	-3	9	0.281
22	20	2	4	0.200
29	28	1	1	0.036
67	64	3	9	0.141
38	40	-2	4	0.200
55	56	-1	1	0.018
53	48	5	25	0.521
32	30	2	4	0.133
35	42	-7	49	1.167
11	16	-5	25	1.562
8	10	-2	4	0.400
21	14	7	49	3.500
400	400			8.159

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right] = 8.059$$

$$\text{ndf} = (4-1) \times (3-1) = 6$$

Table value for 6 df at 5% level = 12.59

Conclusion:

H_0 is rejected since the calculated value of $\chi^2 >$ the table value of χ^2 .

\therefore There is evidence for a significant association between region and grade hotel.

An insurance company advertises in the press a special pension plan for self-employed persons. The advertisement includes a coupon which enables interested persons to complete and return to the company. The company then posts to the enquiries information about the pension plan. If there is no response from the enquiries to the initial information, a second information pack is sent to the enquiries. If there is still no response, a telephone call is made to the enquiries. Enquirers are divided by the company into three categories: definitely takes on a plan, shows interests in plan, not interested. The company analysed a sample of 200 respondents to the initial advertisement (i.e.), those who returned the coupon. The following data was obtained.

	Responds to I mailing	Responds to II mailing	Telephone call made
Takes out plan	36	24	30
Shows interest	18	16	16
Not interested	6	20	34

Test whether there is any association between response and interest in the pension plan.

Solution:

H_0 : There is no association between response and interest in the pension plan.

H_1 : There is association between response and interest in the pension plan.

Observed Frequency Table

	Responds to I mailing	Responds to II mailing	Telephone call made	Total
Takes out plan	36	24	30	90
Shows interest	18	16	16	50
Not interested	6	20	34	60
Total	60	60	80	200

Expected Frequency Table

	Responds to I mailing	Responds to II mailing	Telephone call made	Total
Takes out plan	$\frac{60 \times 90}{200} = 27$	27	36	90
Shows interest	15	15	20	50
Not interested	18	18	24	60
Total	60	60	80	200

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
36	27	9	81	3.000
24	27	-3	9	0.333
30	36	-6	36	1.000
18	15	3	9	0.600
16	15	1	1	0.067
16	20	-4	16	0.800
6	18	-12	144	8.000
20	18	2	4	0.222
34	24	10	100	4.167
200	200			18.189

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right] = 18.189$$

$$\text{ndf} = (3-1) \times (3-1) = 4$$

Table value of χ^2 at 5% level = 9.49

Conclusion:

H_0 is rejected since the calculated value of $\chi^2 >$ the table value of χ^2 .

Hence there is association between response and interest in the pension plan.

A credit rating agency conducted a survey of customers and analyses them by occupation and credit risk. The results were as follows.

Credit rating	Administrative & clerical	Skilled manual	Semi-skilled & unskilled	Total
High	60	50	10	120
Average	30	20	10	60
Poor	10	10	40	60
Total	100	80	60	240

Test whether there is any association between occupation and credit rating.

Solution:

H_0 : There is no association between occupation and credit rating.

H_1 : There is association between occupation and credit rating.

Expected Frequency Table

Credit rating	Administrative & clerical	Skilled manual	Semi-skilled & unskilled	Total
High	$\frac{100 \times 120}{240} = 50$	40	30	120
Average	25	20	15	60
Poor	25	20	15	60
Total	100	80	60	240

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
60	50	10	100	2.00
50	40	10	100	2.50
10	30	-20	400	13.33
30	25	5	25	1.00
20	20	0	0	0.00
10	15	-5	25	1.67
10	25	-15	225	9.00
10	20	-10	100	5.00
40	15	25	625	41.67
240	240			76.17

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right] = 76.17$$

$$\text{ndf} = (3-1) \times (3-1) = 4$$

Table value of χ^2 for 4 df at 5% level = 9.49

Conclusion:

H_0 is rejected since the calculated value of $\chi^2 >$ table value.

\therefore There is association between occupation and credit rating.